 doi:10.20944/preprints202107.0036.v1

Article

Real-Time Quality Index to Control Data Loss in Real-Life Cardiac Monitoring Applications

Gaël Vila [1,2], Christelle Godin [1,*], Sylvie Charbonnier [2] and Aurélie Campagne [3]

[1] Univ. Grenoble Alpes (F-38000 Grenoble, France) & CEA, LETI, MINATEC Campus (F-38054 Grenoble, France) ; gael.vila@cea.fr, christelle.godin@cea.fr,
[2] Gipsa-Lab, Univ. Grenoble Alpes & CNRS, F-38402 Grenoble, France; sylvie.charbonnier@gipsa-lab.grenoble-inp.fr
[3] LPNC UMR 5105, Univ. Grenoble Alpes & CNRS, F-38040, Grenoble, France; aurelie.campagne@univ-grenoble-alpes.fr
* Correspondence: christelle.godin@cea.fr, +334-38-78-40-67 (C.G.)

Abstract: Wearable cardiac sensors pave the way to advanced cardiac monitoring applications based on heart rate variability (HRV). In real-life settings, heart rate (HR) measurements are subject to motion artifacts that can be timely removed from the recordings. This leads to frequent data loss in the HR signal, especially for commercial devices based on photoplethysmography (PPG). The current study had two main goals: (i) to provide a white-box quality index that estimates the amount of missing samples in any piece of HR signal; and (ii) to quantify the impact of data loss on feature extraction in a PPG-based HR signal. This was done by comparing real-life recordings from commercial sensors featuring both PPG (Empatica E4) and ECG (Zephyr BioHarness 3). After an outlier rejection process, our quality index was used to isolate portions of ECG-based HR signal that could be used as benchmark, to validate the output of Empatica E4 at the signal level and at the feature level. Our results showed high accuracy for estimating the mean HR, poor accuracy for short-term HRV features and moderate accuracy for longer-term HRV features. Levels of error could be dramatically reduced by using our quality index to identify time windows with few or no missing data.

Keywords: wearable cardiac sensors; electrocardiography; photoplethysmography; heart rate variability; signal quality; real-life measurements.

1. Introduction

1.1. Broad Context

Among the host of wearable sensors implied in the *Quantified Self* nebula [1], physiological sensing stands a promising tool as it may provide objective information over subject's inner states. Widespread applications are to be found in affective computing [2,3], an emergent and multidisciplinary research field that intends to assess and reproduce mental states form their peripheral and behavioral correlates. Wearable physiological sensors also pave the way to smart healthcare. Clinical issues, like chronic disease management or heart event prevention, may soon be achieved outside the hospital thanks to advanced monitoring systems [4,5].

Whatever the end purpose, feature extraction algorithms (that compute some parameters of interest) are usually run on physiological signals to detect or estimate the target state (*e.g.* subject's stress level or risk for heart attack). In this process, cardiac activity through heart rate (HR) provides a major source of information. Advanced cardiac monitoring can be achieved through a set of statistical, frequency or geometrical features which quantify heart rate variability (HRV) [6]. Since they correlate with autonomic nervous activity, these features allow to predict target states or crisis events – such as stress [7], mental load [8], epilepsy [9]. To reliably monitor these HRV features, the ideal cardiac

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 July 2021 doi:10.20944/preprints202107.0036.v1

sensor should be able to timely estimate the heart rate in ambulatory conditions and uncontrolled environments.

1.2. Wearable Technology for Heart Rate Estimation

The gold standard measure to extract HR is electrocardiography (ECG), in which heartbeats are markedly identified by sharp and prominent peaks (the R-waves). Instantaneous heart rates are derived from the interbeat intervals (IBI), *i.e.*, the time differences between successive heartbeats.

Wearable ECG can be achieved through textile electrodes (textrodes, or e-textiles [10]), which have shown performances comparable with the gold-standard Holter technology for heart rate estimation [11]. While Holter-type devices require expert knowledge to set up and become obtrusive in case of physical activity, textrode-based devices have been successfully integrated in chest belts or usual garment (smart T-shirts for instance). This makes e-textiles suitable for the widespread market and paves the way to real-life cardiac monitoring through ECG. To this day, however, they have been mainly commercialized in specialized areas, like exercise monitoring, and have not been embraced by the common people yet.

Meanwhile, wrist-worn sensors with abilities in cardiac monitoring (*e.g.* smart watches or smart wristbands) are now used by increasing numbers of users, worldwide and daily. Most of these sensors enable photoplethysmography (PPG), a standard measure of oxygen saturation in blood [12]. PPG provides an oscillatory signal made of pulse waves, which can be used to identify heartbeats and estimate the IBI.

In controlled environments, the IBI signals extracted from PPG and ECG correlate well under static conditions [13,14]. However, HRV parameters have also shown substantial deviations between PPG and ECG under the influence of moderate mental effort or wrist activity [14]. Compared to the sharp R-waves in ECG, indeed, the oscillatory nature of PPG waveforms makes heartbeats difficult to locate in time with high accuracy. Moreover, the varying pulse transit time between heart and the wrist may lead to a different estimate of the IBI at these two locations, depending on dynamic physiological factors like blood pressure [15]. To account for such differences, HRV is sometimes called *pulse rate variability* when measured through PPG.

In ambulatory settings, an additional source of error in IBI estimation arises from motion artifacts, which especially affect PPG measurements. Over the past years, artifact identification and removal has been an active field of research on cardiac signals from both ECG and PPG [16,17]. A more comprehensive approach consists in computing a signal quality index (SQI) on the cardiac recordings. This index can be used to timely skip heartbeat detection when the raw signal is flawed.

1.3. Academic Research on Signal Quality

Regarding ECG, statistical or frequency parameters can be extracted from the raw signal to assess its quality or infer a signal-to-noise ratio [18]. These ECG features can be combined using machine learning techniques. Using support vector machines, the SQI presented in [19] obtained an accuracy of 93% in classifying 10-second segments of arrhythmic recording as clinically acceptable or not, with manual annotation as a reference. Regarding PPG, the pulsatile waveform can be compared to a template based on expert knowledge and the surrounding beats. This technique allowed [20] to reach 95% accuracy while predicting acceptable pulses against expert annotation, and [21] to reach a true positive rate of almost 100% in beat detection by setting a threshold on their own SQI – at the price of missing one beat over 10.

An alternative approach was presented in [22], where HRV was directly used to compute an SQI over 30-second segments of IBI signal from ECG recordings in ambulatory settings. Wavelet entropy was extracted from high-frequency ranges and fed in support vector machines, with expert annotation as desired output. The final algorithm reached 94% in accuracy on the test set. An extension of this approach is proposed in [23], where

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 July 2021 doi:10.20944/preprints202107.0036.v1

the SQI (developed on both ECG and PPG recordings), was based on a mixed use of the IBI estimate (acceptable range and variations) and the raw sensor data (template matching strategy).

In each experiment, however, the data were collected in laboratory or clinical settings, using standard instrumentation (like Holter monitors or finger-worn PPG devices), presumably set up by some expert personnel. The proposed SQIs still need validation in real-life settings, where signals are much more exposed to motion artifacts, and collected from commercial equipment installed by the users themselves. The quality index presented in this article was not designed to outreach these SQIs in accuracy for detecting flawed signal segments on standardized data. It aims to provide an easily understandable criterion to select reliable time widows in real-life IBI signals from wearable sensors, assuming the beat detection is skipped in case of movement artifacts.

1.4. Advanced Cardiac Monitoring with Commercial Sensors

Indeed, some commercial devices compute their own SQIs to control the reliability of their heart rate outputs in case of noisy data. This is the case, for example, with chest belt Zephyr BioHarness 3 that produces a confidence level on its ECG-based HR indicator (see below). Another well-known device in the sensor market is the smart wristband Empatica E4, based on PPG. To timely account for data quality, the device returns an IBI signal only when PPG waveforms are considered consistent enough by the heartbeat detection algorithm [24]. In other words: the IBI output is bound to some implicit, binary SQI whose computation has not been disclosed yet by the company.

The E4 wristband has already undergone several validation studies, which focused on the signal level (accuracy of IBI estimation) and/or the feature level (accuracy of the feature extraction process). For example, [25] estimated the number of missing beats in the IBI signal, relatively to standard ECG laboratory instrumentation; and compared statistical and frequency features from both devices. In the same vein, [26] compared IBI values and cardiac features from E4 with an ambulatory ECG device (VU-AMS) in clinical settings ; and [27] collected cardiac features from E4 and another wearable ECG instrument (MindWare Mobile Device).

These studies are consistent in their findings. At the signal level, the IBI estimate from E4 is well correlated with the IBI from ECG, with better results in resting than active conditions. In a given time window, however, the proportion of missing samples could reach 57% at rest and 99% during a talk [25], due to the heartbeat selection algorithm embedded in E4. At the feature level, the mean heart rate over a given time window is estimated with high accuracy; but all features reflecting HRV show significant correlations [27] together with significant differences [26] with the ECG-based data.

This deviation from the gold-standard reflects the well-known limitations of PPG in assessing HRV. That said, it also comes along with considerable loss of data on time intervals where the PPG is likely to be flawed by motion artifacts. Hence, it is still unclear whether this substandard cardiac monitoring with Empatica E4 comes from (i) permanent limitations of PPG or from (ii) transient data loss due to motion artifacts. In case (ii), there would be room for a quality management strategy in the feature extraction process by skipping feature computation when signal quality is too low.

Finally, none of these validation studies has been conducted in real-life settings, where motion artifacts are frequent and measurement devices are set up by users themselves. In such conditions, a methodological obstacle has to be overcome: there is no gold-standard measurement to compare with the validated device. Since any wearable sensor (including ECG) is exposed to motion artifacts, there is a need to select reliable segments of IBI data to act as a benchmark. In that perspective, expert annotation cannot make a sustainable strategy; and academic SQIs still need validation on real-life data. Since most commercial SQIs are black boxes, they provide both help and burden to the researcher when it comes to select a commercial device for advanced cardiac monitoring. Our work

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 July 2021 doi:10.20944/preprints202107.0036.v1

attempted to address this methodological issue in a field validation procedure for Empatica E4.

1.5. Outline of the Current Study

This article proposes the following contributions to current research on real-life cardiac monitoring systems.

i. A white-box SQI that quantifies the data loss in any IBI signal;
ii. A straightforward criterion to select reliable IBI segments on the field;
iii. Validation results for a wrist-worn sensor (Empatica E4) in real-life settings;
iv. Improvement reports when accounting for data loss in the feature extraction process.

In a first step, our index of data loss (the Lack Index) is developed using three properties of the IBI signal: its acceptable range, its acceptable variability and the sum of its acceptable sample values. This SQI was used to identify time intervals where a wearable ECG device (Zephyr BioHarness 3) could be used as reference for heart rate estimation. In a second step, validity of sensor Empatica E4 a surrogate of wearable ECG is addressed at the signal level and at the feature level, in real-life data acquired by non-expert users. The validation method is similar to [25] : at the signal level, a beat-to-beat analysis is run to compare the IBI signals from both sensors. At the feature level, statistical and frequency parameters are compared between both signals. In this process, the Lack Index allowed us to select time windows in which large error rates were less likely to be encountered.

2. Materials and Methods

This study relies on a database acquired on several subjects, recorded during daytime for a whole working week. The experimental protocol and resulting dataset is introduced in the next paragraphs (*Materials*); then the main data processing and methods for elaborating the SQI is presented in detail (*Methods*); along with the validation procedure to estimate PPG data quality.

2.1. Materials : Experimental Protocol and Cardiac Sensors

2.1.1. Recruitment procedure

Three male participants (aged 25, 27 and 33) agreed to wear a set of commercial sensors during a whole working week. None of them had history of neurological disease, or followed any treatment susceptible to alter their cerebral or neurological functions.

Before the experiment, each participant received a 2-hours briefing, and a detailed manuscript explaining both the protocol and proper use of the sensors, which had to be taken home for the week. Sensors would be equipped, taken off and recharged everyday by participant themselves according to experimenter's instructions, and then brought back to the lab for the final debriefing.

All participants signed a written consent at the end of the initial briefing, and received financial compensation after the equipment was returned. Experimental protocol was approved by the Ethics Committee in Non-Interventional Research (CERNI) related to COMUE Univ. Grenoble-Alpes, and conducted in accordance with the Declaration of Helsinki.

2.1.2. Wearable Sensors and Cardiovascular Signals

Two commercial wearable sensors were used to monitor cardiovascular activity: a chest belt and a smart wristband.

The chest belt Zehyr BioHarness 3 allows continuous electrocardiography (**ECG**$_{bh}$, sampled at 250Hz) by means of a couple of textrodes embedded on a chest strap. Acquisitions are recorded in a compact module clamped on the strap, which power supply and memory exceed a full recording day. The device automatically computes two kinds of heart rate estimators. The first one is a standard tacogram, sampled every time a new heart

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 July 2021 doi:10.20944/preprints202107.0036.v1

beat is detected on signal ECG_{bh}. The interbeat intervals (**IBI_{bh}**) are computed by differentiating all timestamps. The second one is a custom heart rate approximation (**HR_{bh}**), which is computed from the surrounding 15 seconds of signal IBI_{bh} and sampled every second. The reliability of this HR signal was asserted on the field in [28], in low- and high-physical activity conditions. As stated before, BioHarness 3 also returns its own (proprietary) SQI: for each sample HR_{bh}, a confidence level (**C_{hr}**) is returned as a percentage of maximum confidence (hence varying between 0% and 100%). According to BioHarness user manuals, this confidence level C_{hr} is computed from features expressing the maximum level of the raw signal ECG_{bh} and its signal-to-noise ratio. The device additionally provides: (i) 3-axis accelerometer data, which will be used below to compute the amount of body movement, and (ii) a Respiration signal, which will not be considered in the current study.

The smart wristband Empatica E4 provides photoplethysmography with green and red light (**PPG_{e4}**, sampled at 64Hz) and automatically computes the interbeat interval (**IBI_{e4}**). As mentioned earlier, signal IBI_{e4} is sampled when signal PPG_{e4} is considered reliable enough by the heartbeat detection algorithm. E4 additionally provides skin temperature and skin conductance data that will not be considered in the current study. It also provides 3-axis accelerometer data, which will be used below to estimate wrist activity. The device was worn on each subject's non-dominant arm.

Both sensors can be used in local recording or Bluetooth transmission mode. The recording mode was preferred for battery saving and to avoid accidental connection losses. A main challenge with such a protocol choice, is that both sensors had to be synchronized manually.

2.1.3. Sensor Synchronization and Data Collection

As stated before, each sensor was set and started by participants themselves in the morning (*i.e.*, as soon as possible after getting up). After self-calibration of the sensors (*i.e.* one minute after set up), participants were instructed to stand still and perform three successive jumps at one-second intervals. This particular, contrived movement leaves sharp and salient marks on the acceleration signals of each sensor: this allows offline sensor synchronization with a time resolution high enough for our later analyses (around 0.1s). The same procedure was applied before sensors unset at the end of the day (*i.e.*, as late as possible before sleep), to account for time drift between both sensor clocks.

Participant's activities and their interactions with the sensors are another paramount information when collecting data in real life during an extended period. Participants performed their daily routine wearing both sensors (office-like work, bike or car driving, leisure and sport training, two of them being high-level athletes). They were also asked during daytime to report any meaningful event by means of a short questionnaire filled on a homemade smartphone application. Moreover, they had to recall their main activities every evening by means of a spreadsheet questionnaire on the provided computer. The day reconstruction method followed the guidelines of [29]. After uploading together sensor data and daily questionnaires to laboratory's own secured network, at the end of each day, they had a short interview with the experimenter by mail or by telephone depending on participant's own agenda.

2.1.4. Preliminary Processing and Data Selection

Offline, all signal processing was performed using MathWorks® Matlab R2017b. Sensor data were synchronized on the basis of the successive jumps (by identifying them and expanding signal timeframes for a perfect match); and examined in the light of our knowledge concerning subject activities during the experiments. Concerning heart rate data, the beat-to-beat analyses to come required perfect sensor synchronization. Therefore, every record in which the three successive jumps were not clearly identifiable on both sensors' acceleration data, twice a day (since participants might have neglected them at one or the other end of the day), were discarded from current analysis.

Subsequently, heart rate signals from both sensors were superposed and visually examined. The HR estimator from BioHarness (HR_{bh}) was inverted to match the values of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 July 2021 doi:10.20944/preprints202107.0036.v1

the IBI estimator (IBI_{bh}). It then appeared that signal IBI_{bh} could not exceed 2^{15} milliseconds in amplitude, which caused an adverse effect: as soon as no R-peak was detected during more than 32 seconds on ECG, IBI_{bh} fell out of sync with the two other signals (HR_{bh}, IBI_{e4}). This happened a couple of times on several day records. To deal with this issue, IBI_{bh} was synchronized back with HR_{bh} manually and piecewise, by correlation maximization on each correct portion of signal. Finally, the wristband signal IBI_{e4} was also synchronized with the chestbelt signal HR_{bh} (by correlation maximization) to account for the mean pulse transit time between heart and the wrist.

During this two-step, high-precision synchronization phase, every recording day in which the proposed method needed further improvements was discarded from the final dataset. This left 11 recording days, *i.e.* 124 hours of multimodal recordings.

To validate the heartbeat detection algorithm of BioHarness (which provides IBI_{bh}) with an academic reference, Pan-Tompkins (1985) analysis was finally run on each ECG recording [30,31]. This provided a 3rd IBI signal hereafter named **IBI_{pt}**. Again, this signal was synchronized with other IBI estimators by correlation maximization.

2.2. Methods: Signal Processing and Quality Estimation

Our final goal was to compare the PPG-based signal IBI_{e4} with the ECG-based signal IBI_{bh}, on time intervals were signal IBI_{bh} could be used as a benchmark. This implied to find such time intervals given that, unlike IBI_{e4}, signals IBI_{bh} and IBI_{pt} were not corrected for flawed heartbeat detection. To this end, a preliminary rejection of outlier samples was performed on each IBI estimator following an automatic process. In each cardiac signal without outliers, the proportion of missing samples was then estimated over 1-minute time windows of data. This proportion of missing samples is the output of our new SQI. To isolate time windows where no sample was missing (*i.e.*, where flawless heartbeat detection could be assumed), a criterion was applied on this SQI. On time windows where signal IBI_{bh} could be used as benchmark, the quality of signal IBI_{e4} was finally assessed at the signal level (by matching the heartbeats from both signals) and the feature level (by extracted cardiac features from both signals).

2.2.1. Outlier Removal

The IBI signal has inner properties that can be used to assess data quality. These properties were used to reject outlier samples on each IBI signal, by applying two successive criteria: a range criterion and a variation criterion. The range criterion consisted in rejecting every sample that expressed an instantaneous heart rate outside the range: [30; 250] bpm (kept wide to account for sport training events). The variation criterion consisted in considering that a value of an IBI sample cannot deviate from the mean of its neighbors by more than 30%. The whole outlier rejection procedure followed the guidelines of [32,33]. Since real-life IBI signals may sometimes be very noisy, rejection from the variation criterion was performed in a recursive fashion. A 5-s moving average was computed at first by focusing on each sample of the IBI signal, successively. Each IBI value deviating from this moving average by more than 30% was temporarily regarded as an outlier. In the following iteration, the moving average was computed on the remaining non-outlier samples. All IBI samples deviating from this new moving average by more than 30% were regarded as the new outliers – and so on. The procedure was run until no more outlier was found, or when the whole process exceeded 20 iterations. Outputs of this two-steps outlier removal process are illustrated by **Figure 1** below.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 July 2021 doi:10.20944/preprints202107.0036.v1

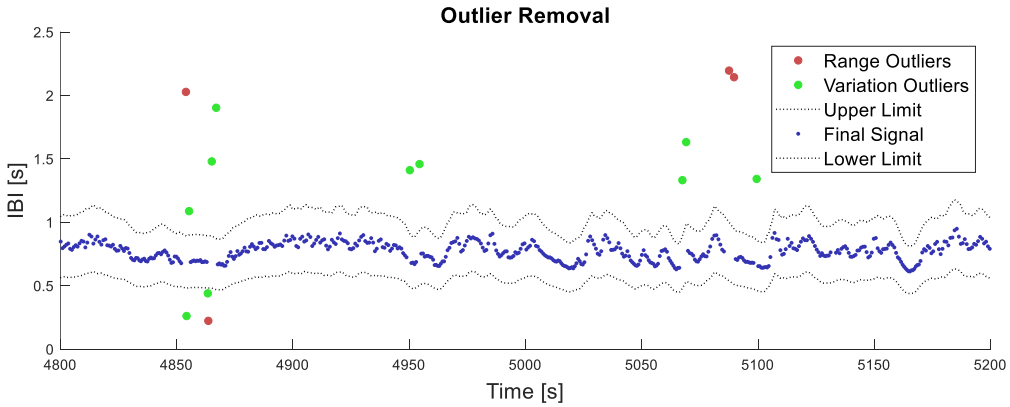


Figure 1. Illustration of the outlier removal process. The red dots represent outliers that were removed in the first place, from the range criterion on IBI values. The green dots represent outliers that have been removed in the second place, from the variation criterion on the IBI signal. The grey dotted-lines represent the final lower and upper limits after algorithm convergence on the variation criterion, *i.e.*, 30% deviation from the moving average. The blue points represent the remaining IBI signal after outlier correction.

2.2.2. The Lack Index

In Empatica E4, IBI samples are provided only when the PPG signal is considered good enough by sensor's own algorithm. This results in a scarcer (yet more trustful) signal than the standard tacogram provided by BioHarness. After the outlier rejection step, however, the IBI samples from BioHarness were also scarcer when ECG signal quality was low. If the heartbeat detection algorithm truly detects heartbeats (this hypothesis will be discussed later), then it can be assumed that sample scarcity (after outlier rejection) reflects the quality of the original IBI signal.

A simple index of "sample scarcity" can be designed by comparing the practical number **N** of IBI samples *actually* found in a given time window, with a theoretical number $\mathbf{N_t}$ of IBI samples that *should* be found in a standard tacogram with no outlier rejection. Let a time window of duration **W**, ranging within [t=0; t=W], in which **N** heart beats are detected at times $\{t_1, t_2 \dots t_N\}$. The corresponding interbeat intervals are computed as: $IBI_n = t_n - t_{n-1}$. In a *raw* IBI signal (*i.e.*, without outlier rejection), all samples add up to the total time duration between the first and the last heart beat used for IBI computation. This can be expressed:

$$\sum_{n=1}^{N} IBI_n = \sum_{n=1}^{N} (t_n - t_{n-1}) = t_N - t_0 \approx W \quad (1)$$

where t_0 is the latest beat detected before the beginning of current time window (t=0).

The mean IBI ($\boldsymbol{\mu}$) in current time window is computed by dividing this sum of IBIs by the number of samples **N** found in this time window. Consequently, this number of samples N can be recovered by dividing $\boldsymbol{\mu}$ by the sum of IBIs. Since this sum of IBIs is close to W, according to Equation (1), one can also compute a theoretical number of samples $\mathbf{N_t}$ that is close to N:

$$N = \frac{\sum_{n=1}^{N} IBI_n}{\mu} \approx \frac{W}{\mu} = N_t \quad (2)$$

In other words: when there is no missing sample in the IBI signal, the actual number of samples **N** that can be *counted* in any time window, can also be *estimated* with a theoretical number $\mathbf{N_t}$ derived from the window size (W) and the mean IBI (μ). When samples are removed from the IBI signal, however, the actual number N should decrease while the theoretical number N_t should remain stable. From the relative difference between these

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 July 2021 doi:10.20944/preprints202107.0036.v1

two numbers, we designed an estimator of sample scarcity, hereafter named the Lack Index (**L**). As shown in Equation (3) below, the Lack Index **L** can also be seen as the relative error between the sum of IBI samples and the time window length :

$$L = \frac{N_t - N}{N_t} = \frac{W/\mu - N}{W/\mu} = \frac{W - \sum_{n=1}^{N} IBI_n}{W} \quad (3)$$

The Lack Index estimates the proportion of missing beats in a given time window. Combined with an outlier rejection process, it evaluates the quality of the original IBI signal. In this study, three Lack Indexes ($\mathbf{L_{e4}}$, $\mathbf{L_{bh}}$ and $\mathbf{L_{pt}}$) were respectively computed out of the three IBI estimators (IBI_{e4}, IBI_{bh} and IBI_{pt}), on non-overlapping successive time windows where W was equal to 60s. Regarding signal IBI_{e4}, the Lack Index was used to quantify the lack of samples in real-life settings; and its impact on signal quality at the feature level. Regarding all IBI signals, the Lack Index was used to isolate time windows which can be considered as *flawless*, i.e. without any missing beat after outlier removal. This could be done by setting a rigorous threshold on its value, as shown in the next paragraph.

2.2.3. Criteria to Select Flawless Time Windows

By definition, a given time window of length W encompasses all heartbeat times {t_1, t_2 ... t_N}, but not t_0 (located before t=0) and t_{N+1} (located after t=W). Therefore, we can set the following inequality:

$$(t_N - t_1) < W < (t_{N+1} - t_0) \quad (4)$$

When the heartbeat detection is *flawless* (i.e. no outlier has been found in the current window), Equation (1) applies for the sum of IBI samples and our Lack index L is framed by two critical values:

$$\frac{t_0 - t_1}{W} < \frac{W - \sum_{n=1}^{N} IBI_n}{W} < \frac{t_{N+1} - t_N}{W}$$
$$\frac{-IBI_1}{W} < L < \frac{IBI_{N+1}}{W} \quad (5)$$

Consequently: if the Lack Index L is superior to IBI_{N+1}/W, then current time window contains flawed beat detections for current IBI estimator. Since there is no confidence that IBI_{N+1} has been well estimated, the minimum IBI value found in the time window is a more secure alternative to prevent adjacent time windows from influencing each other. Therefore, *flawless* time windows (*i.e.*, windows where the IBI signal has no missing sample) can be identified for each of the three IBI estimators (*i.e.* IBI_{e4}, IBI_{bh} and IBI_{pt}) when the following criterion is satisfied:

$$W * L < \min(IBI) \quad (6)$$

Alternatively, BioHarness provides its own SQI: the confidence level $\mathbf{C_{hr}}$, extracted from online properties of the ECG, whose accuracy is not under scope in the current study. This confidence level can also be used to select reliable time windows of cardiac signal from the BioHarness sensor, since it guarantees that heartbeat detection is safe each time its value reaches 100%. Hence, one may identify a *flawless* window of cardiac signal when the minimum value of $\mathbf{C_{hr}}$ over the time window is equal to 100%. This additional criterion can be used to validate the previous one in our attempt to find flawless windows on signals IBI_{bh} and IBI_{pt}:

$$\min(C_{hr}) = 100\% \quad (7)$$

In the second part of this study, time segments of signal IBI_{bh} that were identified as *flawless* according to Equation (6), were used as benchmarks to assess the quality of signal IBI_{e4}. This was done at the signal level and at the feature level through two distinct procedures: a beat-to-beat comparison and feature extraction from both signals.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 July 2021 doi:10.20944/preprints202107.0036.v1

2.2.4. Beat-to-Beat Comparison Between IBI Signals from BioHarness and E4

When both signals are properly synchronized, each beat in IBI_{e4} can be identified to one beat in IBI_{bh}. This was done by splitting the timeframe of signal IBI_{bh} at the middle between each couple of successive heartbeats, defining a set of time intervals surrounding each heartbeat (see **Figure 2** below). Samples of signal IBI_{e4} were enumerated within each of those time intervals: an empty interval means one missed beat and an interval with more than one sample IBI_{e4} means (at least) one over-detected beat. In each time interval where no heartbeat was missing or over-detected, a pair of matching IBI samples was identified between the two signals. The absolute difference was computed between all matching pairs of IBI samples.

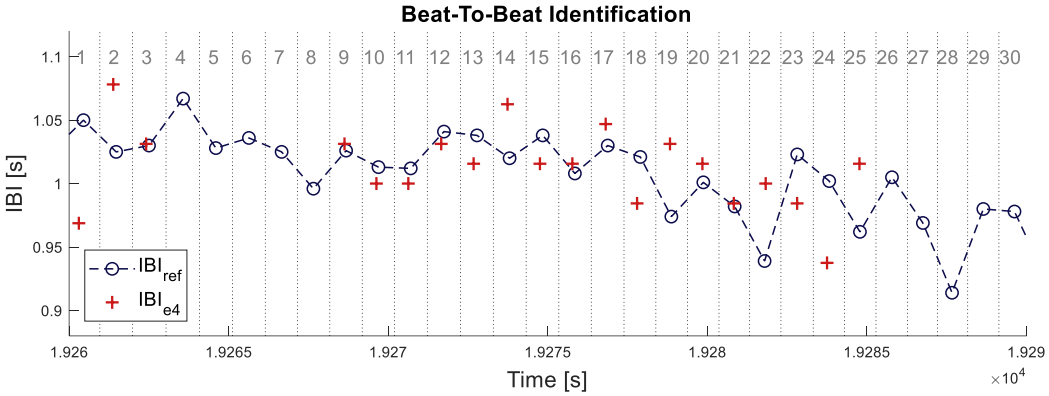


Figure 2. Illustration of beat-to-beat identification procedure in a 30-second time window. Blue circles with dashed line represent the IBI estimate from Zephyr BioHarness 3 (IBI_{bh}). Red crosses represent the IBI estimate from Empatica E4 (IBI_{e4}). Vertical dotted lines delimit the time intervals attached to each IBI sample from BioHarness, which are numbered in grey at the top of the chart. Intervals N°4, 5, 6, 7, 8, 26, 27, 28, 29 and 30 have no red crosses, so that 10 samples (33%) are missing in signal IBI_{e4}. All red crosses here are matched samples, so that no sample is overdetected in signal IBI_{e4}.

This beat-to-beat analysis was run for every time window where signal IBI_{bh} had no missing sample, according to the criterion provided in Equation (6). The following parameters were computed on each single time window: (i) the probability for a sample in signal IBI_{bh} to be missing in signal IBI_{e4} ($\mathbf{p^{miss}}$), (ii) the probability for a sample in IBI_{e4} to be an overdetection ($\mathbf{p^{over}}$), and (iii) the mean absolute difference in IBI value between pairs of matched samples in signals IBI_{bh} and IBI_{e4} ($\boldsymbol{\mu^{diff}}$). These parameters allow to quantify three types of errors that can occur during a beat detection that are, respectively: a false negative, a false positive and a true positive with a wrong value. Let us call $\mathbf{N_{bh}}$ the number of samples for IBI_{bh} in this time window, $\mathbf{N_{e4}}$ the number of samples for IBI_{e4}, n^{miss} the number of missing samples in IBI_{e4}, n^{over} the number of over-detections and $\mathcal{M}$ the set of matched samples between the two signals. The three previous error rates were computed this way:

$$p^{miss} = \frac{n^{miss}}{N_{bh}}, \qquad p^{over} = \frac{n^{over}}{N_{e4}},$$
$$\mu^{diff} = \frac{1}{card(\mathcal{M})} \sum_{n \in \mathcal{M}} |IBI_{e4}(n) - IBI_{bh}(n)| \tag{8}$$

2.2.5. Feature Extraction

In advanced cardiac monitoring applications, descriptive features are usually computed over a full time window. For example, the mean interbeat interval ($\boldsymbol{\mu}$) and their standard deviation ($\boldsymbol{\sigma}$, also called SDNN in current literature) are frequently used in the assessment of mental states. Some features are especially designed to describe HRV: for example, the root mean square of successive differences in the IBI (**rmssd**), and the low-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 July 2021 doi:10.20944/preprints202107.0036.v1

and high-frequency components in its power spectrum (respectively **lf** and **hf**, corresponding to the normalized power in frequency ranges [0.04 0.15[Hz and [0.15; 0.50[Hz).

These five cardiac features were extracted and compared between signals IBI_{bh} and IBI_{e4}, over time intervals where signal IBI_{bh} showed no missing sample (*i.e.*, satisfied the criterion set in Equation (6)). The two frequency features were computed by integrating a Lomb-Scargle periodogram [34], a frequency analysis technique adapted to non-evenly sampled signals. The time window length W=60s, was set to ensure adequate resolution in the lower frequencies.

To assess feature extraction from IBI_{e4} with IBI_{bh} as benchmark, the absolute error rate (**E**) between both estimates was computed for each feature f, over each time window.

$$f \in \{\mu, \sigma, rmssd, lf, hf\}, \qquad E^f = \left|\frac{f_{e4} - f_{bh}}{f_{bh}}\right| \tag{9}$$

This error rate was used to answer two additional questions: (i) What levels of error could be expected for a given cardiac feature when extracted from signal IBI_{e4}? (ii) To what extent could the Lack Index (L_{e4}) be used to reduce this error? Since the Lack Index L_{e4} estimates the proportion of missing samples in a time window, question (ii) tests the hypothesis that lack of samples is a significant factor in misestimating a feature from signal IBI_{e4}. If so, one could set a threshold on L_{e4} to timely control the risk of error when extracting a cardiac feature from real-life IBI recordings.

In practice, however, samples could be extremely rare in signal IBI_{e4} when the heartbeat detection algorithm was unsuccessful over large time intervals. In such contexts, it may be meaningless to compute some of the previous features since the IBI segment contains too little information. Therefore, some conditions were set to compute a given feature (and the corresponding error rate) only when the number of samples $\mathbf{N}_{e4}$ in signal IBI_{e4} was sufficient over a 60s time window:

- μ was computed when Ne4 ≥ 1 sample;
- σ was computed when Ne4 ≥ 2 samples;
- rmssd was computed when Ne4 ≥ 2 successive samples;
- lf and hf were computed when Ne4 ≥ 18 samples.

The minimum number of samples needed to estimate the frequency features depends on the range and resolution of the Lomb-Scargle periodogram, which is different for each window of unevenly sampled IBI signal [35]. To avoid setting rules that are outside the scope of this study, the minimum N_{e4} was thus derived from the Shannon-Nyquist theorem, assuming that power spectrum should spread (at least) beyond the critical frequency of 0.15Hz (hence: $N_{e4} \geq 60_{[s]} * 2 * 0.15_{[Hz]} = 18$).

2.2.6. Activity Level Monitoring

Since previous error rates were computed under ambulatory conditions, the amount of body movement stands as a major feature to monitor in order to explain potentially low performances for each IBI estimator. Using data from its 3-axis accelerometer, the BioHarness sensor computes and returns an estimate of Activity which, according to device documentation, is derived from the Euclidean norm on the 3 bandpass-filtered acceleration components. Although this parameter might accurately reflect the amount of chest movement, there is no equivalent indicator for the wrist-worn sensor E4. A common Activity level estimator was thus computed from the raw acceleration data of both sensors, to quantify body movements on both locations on comparable magnitude scales. Each component was first bandpass-filtered in the frequency range [0.1; 10] Hz, with a digital second-order Butterworth filter, to account for non-human artefacts and the low-frequency contributions of gravity. According to previous studies in human movement quantification [36,37], the sum of each component's average signal magnitude area correlates well with energy expenditure and allows to distinguish between rest and active periods. This can be expressed:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 July 2021 doi:10.20944/preprints202107.0036.v1

$$A = \sum_{i \in 1..3} A_i \,, \qquad A_i = \frac{1}{W} \int_{t=0}^{W} |a_i(t)| dt \tag{10}$$

where a_i is one of the three bandpass-filtered acceleration components, and [0, W] delimits a given time window. Activity **A** was thus computed for both sensors to provide one Activity estimate for each body location: $\mathbf{A_{bh}}$ for the chest and $\mathbf{A_{e4}}$ for the wrist. Over the set of all available 60-second time windows in the database, our Activity estimate A_{bh} got a Spearman correlation coefficient of 0.91 with the mean Activity estimate from BioHarness.

3. Results

This section presents validation results on sensor Empatica E4 through the three stages of our validation method: (i) selection of time intervals with no missing sample; then (ii) validation of estimator IBI_{e4} against estimator IBI_{bh} at the signal level and (iii) at the feature level.

3.1. Time Windows with Flawless Heartbeat Detection

For each of the three IBI estimators (IBI_{bh}, IBI_{pt}, IBI_{e4}), the criterion shown in Equation (6) was applied to isolate 60s time windows where no sample was missing. The criterion based on BioHarness native SQI (Equation (7)) was also applied as a reference. Following each criterion, the size of the selected subset of time windows is displayed in **Table 1** below, as a percentage of the initial database. It represents the probability of selecting a random time window if either the BioHarness SQI ($\mathbf{C_{bh}}$) or the Lack index ($\mathbf{L_{bh}}$, L_{pt}, L_{e4}) had been used to select flawless time windows.

Each subset of time windows also contained various amounts of chest movement ($\mathbf{A_{bh}}$) and wrist movement ($\mathbf{A_{e4}}$). Within each subset, the range of these indicators reflects the amount of body movement that can be monitored when the corresponding criterion is satisfied. The two last lines of **Table 1** thus displays the maximum A_{bh} and maximum A_{e4} over the time windows selected following each criterion. A lower figure means that fewer degrees of physical activity could be monitored through the corresponding signal without losing IBI samples.

Table 1. Proportion of selected 60-second time windows over the whole dataset (upper line), and maximum Activity level in chest (A_{bh}) and wrist (A_{e4}) in the selected time windows (lower lines). Over the whole dataset (*i.e.*, without window selection), the maximum level for A_{bh} was 382 mG and the maximum level for A_{e4} was 825 mG. Column N°2 corresponds to window selection from BioHarness SQI: C_{hr} (Equation (7)). Columns N°3, 4, 5 correspond to window selection from the Lack Index of each IBI signal: L_{bh}, L_{pt} and L_{e4} (Equation (6)).

Criterion	$\mathbf{C_{hr}}$	$\mathbf{L_{bh}}$	$\mathbf{L_{pt}}$	$\mathbf{L_{e4}}$
Selected windows	51.4%	52.4%	47.6%	0.6%
Maximum A_{bh}	137 mG	137 mG	137 mG	7.05 mG
Maximum A_{e4}	415 mG	415 mG	415 mG	19.1 mG

In **Table 1** (line 2, columns 2, 3 and 4), the number of selected windows is similar for the three cardiac signals from BioHarness (HR_{bh}, IBI_{bh} and IBI_{pt}): about half the original database was selected according to each SQI (resp. C_{hr}, L_{bh} and L_{pt}). Three conclusions can be drawn from these three figures. First, no cardiac sensor is immune to real-life artefacts, since half the database showed at least one missing beat after outlier rejection. Second, BioHarness algorithm for heartbeat detection (column 3) performed as well as the academic reference (Pan-Tompkins', column 4) on the same ECG recordings (with resp. 52% and 48% selected time windows). Third, the two criteria used for columns 2 and 3 validated each other in assessing BioHarness cardiac signals. Indeed, the proportions of selected time windows are close to each other (resp. 51.4% and 52.4%); and 91% of the time windows that satisfied Equation (6) (based on the Lack Index L_{bh}) also satisfied Equation

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 July 2021 doi:10.20944/preprints202107.0036.v1

(7) (based on the Confidence Level C_{hr}). The native BioHarness SQI thus validated the ability of the Lack Index for selecting IBI segments with good signal quality.

Compared to the three cardiac signals from BioHarness, however, our criterion selected a very small subset of time windows showing no data loss for signal IBI_{e4} (0.6% of the original database). This significant shrinkage of the subset size (line 2), comes with a significant drop in the amounts of body movement that can be monitored in this subset (lines 3 and 4). Indeed, the maximum levels of chest activity (A_{bh}) and wrist activity (A_{e4}) are identical across columns 2, 3 and 4 (for sensor BioHarness), up to 36% (for chest) and 50% (for wrist) of the maximum activity level in the original database. In column 5 (sensor E4), though, the maximum levels of chest and wrist activity are 20 times lower, down to 1.8% (for chest) and 2.3% (for wrist) of the maximum activity levels in the database.

In a nutshell, **Table 1** shows that time windows where signal IBI_{e4} can be considered *flawless* are very rare in real-life recordings. As expected, the quality of data from sensor Empatica E4 is strongly related to the amount of body movement. The next paragraphs address this issue more accurately, by validating the IBI estimate of Empatica E4 at the signal and the feature level.

3.2. Characterization of Empatica's IBI Estimate at the Signal Level

In the next step, all time windows where signal IBI_{bh} showed missing samples (according to Equation (6)) were removed from the original dataset. This left 3370 time windows (56 hours of recordings) where signal IBI_{bh} could be used as benchmark to validate signal IBI_{e4}. Following the heartbeat-to-heartbeat analysis proposed in section 2.2, we used this new dataset to address the following questions:

- What kinds of error can be expected in signal IBIe4 under real-life conditions (paragraph 3.2.1);
- To what extent are these errors related to wrist movements (paragraph 3.2.2).

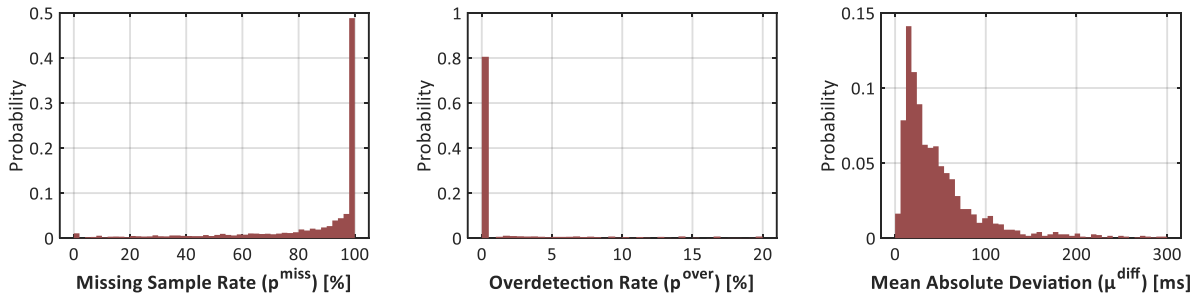
We finally used our results to show that the Lack Index accurately estimates the proportion of missing beats over any time window (paragraph 3.2.3).

3.2.1. Error Rates over All Time Windows

As introduced in section 2.2, three parameters have been computed to qualify the estimation of the IBI by Empatica E4: the proportion of missing samples $\mathbf{p^{miss}}$, the proportion of over detected beats $\mathbf{p^{over}}$, and the mean absolute deviation between matching IBIs $\boldsymbol{\mu^{diff}}$. The repartition of these error rates is pictured with histograms in **Figure 3** below.

Results on the missed sample rate $\mathbf{p^{miss}}$ confirm that signal IBI_{e4} was very scarce over the time windows where signal IBI_{bh} could be used as benchmark. On plot (a), half the time windows have a p^{miss} over 98%. Actually, 42% of the time windows did not show any sample of signal IBI_{e4}. In 90% of the time windows, the heartbeat detection algorithm missed at least one beat over two (*i.e.*, the 1st decile in p^{miss} is close to 50%).

On plot (b), however, one may notice that 81% of the time windows show no over-detected heartbeat (p^{over}=0%). Overdetection is thus a marginal phenomenon in signal IBI_{e4}. In plot (c), one may also notice that deviations between matched samples of signals IBI_{bh} and IBI_{e4} are typically low: 90% of the time windows show a mean deviation (μ^{diff}) below 130ms, which represents roughly 15% of a typical IBI value.



(**a**) (**b**) (**c**)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 July 2021 doi:10.20944/preprints202107.0036.v1

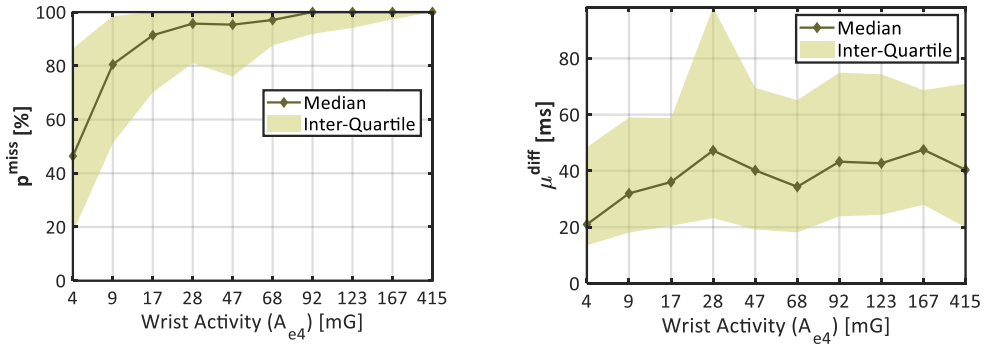
Figure 3. Histograms on three error rates to validate signal IBI_{e4} with signal IBI_{bh} as benchmark. (a) Proportion of missing samples (p^{miss}) in signal IBI_{e4}; (b) Proportion of excess samples (p^{over})in signal IBI_{e4}; (c) Mean Absolute Difference (μ^{diff}) between matching pairs of samples between IBI_{e4} and IBI_{bh}.

These results show that in the estimation of IBI with Empatica E4, the risk of false positive (*i.e.*, an overdetected beat) is very low and the estimation accuracy is typically high. That said, plot (c) also demonstrates that such an estimation is not perfect: the mean value of μ^{diff} over the dataset is 67ms, which represents roughly 7% of a typical IBI. Together with the high rates of missed beats, this should impact the quality of feature extraction in advanced cardiac monitoring applications.

3.2.2. Impact of Wrist Activity

One may still argue that histograms of **Figure 3** include all amounts of body movements, while PPG measurements are known to behave well under conditions of low physical activity. In this paragraph, we measure the same error rates at low and high amounts of wrist movement.

To distinguish several amounts of wrist movement, the dataset was divided in ten subsets of time windows corresponding to each decile of wrist activity $\mathbf{A_{e4}}$ (the lowest 10% make one subset, the next 10% make another subset, *etc.*). In each of these subsets, the median and the interquartile range were computed for the two error rates $\mathbf{p^{miss}}$ and $\boldsymbol{\mu^{diff}}$ (p^{over} was not considered since it was typically negligible). The results are displayed both charts of **Figure 4**.



(**a**) (**b**)

Figure 4. Statistics on IBI estimation quality through distinct levels of wrist activity. (a) Proportion of missed beats (p^{miss}) against wrist activity (A_{e4}); (b) Misestimation degree (μ^{diff}) against wrist activity. The X-Axis represents each decile of A_{e4} in the dataset. Plain lines stand for the median error at each decile of activity; and colored areas materialize the interquartile range.

On chart (a), the median (thick line) and the interquartile range (colored area) of the proportion of missing beats (p^{miss}) together show a steep rise across the lower deciles of wrist activity (A_{e4}<17mG), and remain stable afterwards ($p^{miss} \approx 100\%$). According to Student's two-sample t-test, there was a highly significant difference in p^{miss} between the two first deciles of wrist activity ($p < 0.001$, 672 degrees of freedom). These results confirm that wrist movements strongly increased the probability of missing a heartbeat, even at the lowest rates (median $p^{miss} \approx 80\%$ in the second decile of A_{e4}).

On chart (b), the mean absolute deviation (μ^{diff}) shows a comparable trend: the median error increases across the first deciles of wrist activity (from 21ms at 4mG to 36ms at 17mG); and tends to remain stable afterwards. A one-way ANOVA across all deciles of A_{e4} showed a significant impact of this factor on parameter μ^{diff} ($p<0.001$, 1952 degrees of freedom). However, the median curve remains encased in parameter's variability (colored area); and there is no significant difference between the two first deciles of wrist activity ($p = 0.283$, 563 degrees of freedom). Compared to p^{miss}, the mean deviation μ^{diff} was thus only mildly affected by the amount of wrist movement.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 July 2021 doi:10.20944/preprints202107.0036.v1

During the final step of this study (see section 143.3), the 1rst decile of wrist activity was used to delimit a subset of time windows showing better IBI estimation, since p^{miss} and μ^{diff} are both lower than in the rest of the dataset.

3.2.3. Validation of the Lack Index

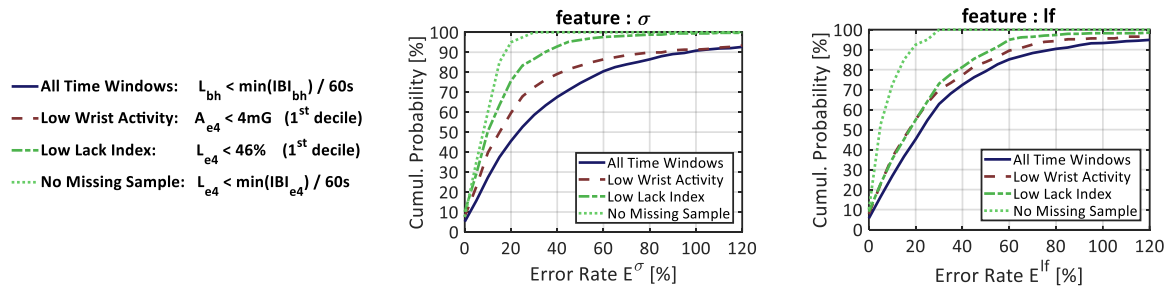
In the current study, the Lack Index ($\mathbf{L_{e4}}$) and parameter $\mathbf{p^{miss}}$ are two distinct measures of the proportion of missing beats in a given time window. Over all time windows where signal IBI_{bh} could be used as benchmark, the linear correlation coefficient found between L_{e4} and p^{miss} was 0.999. This suggests that, despites the underlying hypotheses, our Lack index precisely estimates the actual proportion of missing samples in a time window. The mean absolute deviation between the two indicators is 0.6%. Since an extremely low IBI (*e.g.* 0.5s) covers 0.8% of a 60s time window, the minimum IBI divided by window length (60s) was a relevant upper limit on L_{e4} to claim that a time window had no missing sample (p^{miss} = 0%). This validates the criterion proposed in Equation (6).

3.3. Validation of Empatica E4 at the Feature Level

Finally, the reliability of signal IBI_{e4} for advanced cardiac monitoring was addressed by extracting the five features introduced in paragraph 2.2.5. For each feature $\mathbf{f} \in \{\boldsymbol{\mu}, \boldsymbol{\sigma}, \mathbf{rmssd}, \mathbf{lf}, \mathbf{hf}\}$, the corresponding error rate $\mathbf{E^f}$ between signals IBI_{e4} and IBI_{bh} was computed following Equation **Error! Reference source not found.**.

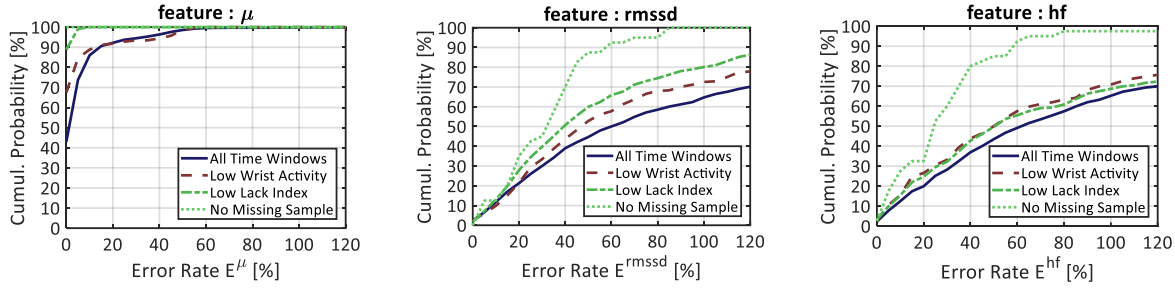
These error rates were first computed over the 3370 time windows where signal IBI_{bh} could be used as benchmark (*i.e.*, where the Lack index $\mathbf{L_{bh}}$ satisfied the criterion defined in Equation (6)). The results are displayed on **Figure 5** below. On each of the five panels, the plain blue line materializes the cumulative probability of error for a given feature *f*. This function corresponds to the probability for error E^f to be less than or equal to a given value. The median error is obtained when the curve reaches 50% in ordinate. The full curve can be interpreted like a ROC curve: when it rises steeply from 0% to 100%, then lower error rates come with higher probabilities, which means that feature extraction from signal IBI_{e4} is more reliable.

When features are blindly computed over all time windows (plain blue line) the median errors are, in ascendant order: **μ**: 3% – **σ**: 25% – **lf**: 25% – **rmssd**: 62% – **hf**: 63%. The mean IBI (**μ**) by far is the more accurately estimated: almost 90% of the time windows have an error rate below 10%. The curve of each other temporal feature resembles the curve of a frequency feature: **σ** and **lf** on the one hand, **rmssd** and **hf** on the other hand. Such similarities, which justified the positioning of the features on **Figure 5**, are not surprising since these couples of features are known to correlate each other in long recordings [38]. In panels (b) and (c), the long-term HRV features (**σ**, **lf**) show together lower error rates than the short-term HRV features (**rmssd**, **hf**), but also higher rates than the mean IBI (**μ**).



(**a**) (**b**) (**c**)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 July 2021 doi:10.20944/preprints202107.0036.v1



(**d**) (**e**) (**f**)

Figure 5. Cumulative probability of error for five features extracted from signal IBI_{e4}, with signal IBI_{bh} as benchmark. (a) Main legend box; (b) Error on the standard deviation of the IBI (**σ**); (c) Error on the normalized power in the low-frequency range (**lf**); (d) Error on the mean IBI (**μ**); (e) Error in the root mean square of successive differences in the IBI (**rmssd**); (f) Error on the normalized power in the high-frequency range (**lf**).
Each feature was computed on 4 different datasets: (i) over all time windows (blue plain line); (ii) over time windows that belong to the 1st decile of wrist activity $\mathbf{A}_{e4}$ (red dashed line); (iii) over time windows that belong to the 1st decile in the proportion of missing samples, estimated by the Lack index $\mathbf{L}_{e4}$ (green dashed-dotted line); (iv) over time windows showing no missing sample according to the Lack Index $\mathbf{L}_{e4}$ (green dotted line).

To reduce such error rates and ensure a reliable feature extraction, one may set a criterion to select time windows where signal IBI_{e4} should be more reliable. As stated in a previous paragraph (3.2.2), for example, the first decile of wrist activity (A_{e4}) defines a subset of time windows where signal IBI_{e4} shows lower error rates at the signal level. The cumulative probability of error for this "Low Wrist Activity" subset is materialized by the red dashed lines in all panels of **Figure 5**.

Since we have designed a SQI (the Lack Index) that estimates the proportion of missing samples in a time window, we also questioned the influence of this parameter on feature estimation quality. To mirror the previous segmentation over wrist activity (A_{e4}), the first decile of the Lack Index (L_{e4}) made another subset of time windows with "Low Lack index". The cumulative probabilities of error in this subset are provided by the green, dash-dotted line on each chart of **Figure 5**.

Finally, a third subset of data was made from the time windows where signal IBI_{e4} had no missing sample, according to Equation (6). This more stringent criterion was designed to suppress the effect of data loss during feature extraction. On each chart of **Figure 5**, the cumulative probabilities of error in this subset are represented by the green dotted line.

As expected, the "Low Wrist Activity" data (red dashed lines) yielded lower error rates than the full dataset. The "Low Lack Index" subset, however, yields better results for every feature except **hf**. Regarding signal quality for feature extraction, therefore, the Lack Index seems to have greater discriminative power than physical activity estimates.

Indeed, the "No Missing Sample" criterion (based on the Lack Index) performed by far the best segment selection. On chart (d), the green dotted curve is merged with the top of the graph: this means that all time windows showed error rates close to 0%. On the other charts, however, the feature estimation is still imperfect., the median error rates being: **σ**: 10% – **lf**: 6% – **rmssd**: 34%– **hf**: 27%. If the long-term HRV features (**σ**, **lf**) now display low error rates, this error remains high for the short-term HRV features (**rmssd**, **hf**) even when no sample is missing in the time window.

Beyond its impact on feature extraction, the low number of available samples in IBI_{e4} also reduced the number of time windows available for feature computation. Over all time windows where IBI_{bh} could be used as reference, the mean IBI (μ) could be extracted 58% of the time (since at least 1 sample was required); features **σ** and **rmssd**, 51% of the time (since 2 samples were required); and features **lf** and **hf**, 23% of the time (since 18 samples were required).

4. Discussion

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 July 2021 doi:10.20944/preprints202107.0036.v1

The contributions of this study are twofold: (i) a validation experiment of Empatica E4 on the field, (ii) a white-box SQI and a methodological framework to qualify cardiac signals on the field.

4.1. Field Validation of Empatica E4

Technically speaking, Empatica's built-in algorithm (that returns an IBI estimation only under conditions of satisfactory PPG quality) is working quite well. In our dataset, over-detection of heartbeats remained a marginal phenomenon. The IBI estimation error was typically low and this misestimation degree increased only moderately with physical activity (**Figure 4** (b)). In other words, Empatica's heartbeat detection algorithm is both specific (few false positives) and accurate (true IBI values). In time windows where all heartbeats were detected by Empatica's algorithm, perfect computation of the mean IBI (**Figure 5** (d)) demonstrates that signal IBI_{e4} was not biased. These results confirm the potential of PPG measurements for monitoring instantaneous pulse rates, which is already intended by a number of wrist-worn cardiac sensors in commercial settings.

At the signal level, such an accuracy is achieved at the price of a very scarce estimation, as shown by parameter p^{miss} in **Figure 3** (a). This lack of available samples measured in real-life settings supports the findings of earlier laboratory studies [25]. At the feature level, our study demonstrated that data loss is a major source of error when trying to extract HR and HRV features: reducing the proportion of missing samples also reduces the error rates, as seen in every chart of **Figure 5**. This factor is bounded to (but not entirely explained by) the amount of wrist activity (see **Figure 4** (a)).

4.2. Assets and Drawbacks of the Lack Index

The Lack Index was designed to deal with data loss in real-life recordings. According to paragraph 3.2.3, it precisely estimates the proportion of missing samples in any time window of IBI signal. It also comes with a straightforward criterion to select time windows in which no sample is missing (see Equation (6)).

That said, the Lack Index assumes that all IBI samples in a time window are valid heart rate data. To be used as an SQI, it should be combined with an outlier rejection procedure to ensure that IBI samples have all the properties of a valid cardiac signal. The procedure implemented in Section II.B is a perfectible example of such a process: since it relied on neighboring data, it was unable to identify a too long succession of invalid cardiac measurements. As presented in this study, the Lack Index thus assumed that sensor's heartbeat detection algorithm was efficient enough to guarantee that non-outlier samples reflect true IBI data. This assumption was verified for sensors Zephyr BioHarness 3 and Empatica E4, since they validated each other in the second part of this study.

Along with the Lack Index, we hence proposed and validated a white-box method to qualify a heart rate signal in real-life settings. It consists in two main steps: (i) following Equation (6), select time intervals where a reference measurement (*e.g.*, wearable ECG) shows no missing sample after outlier rejection; and then (ii) run a validation study with the previous IBI segments as benchmark.

While selecting IBI segments with no missing samples, the results of **Table 1** suggested that for Empatica E4, real-life cardiac monitoring applications should always rely on incomplete segments of IBI signal (no sample was missing in only 0.6% of the dataset). To implement a sustainable feature extraction process, a compromise should be brought on an "acceptable" number of missed beats for each individual feature. The Lack Index was designed to deal with such an issue.

4.3. Potential in Advanced Cardiac Monitoring Applications

On **Figure 5**, indeed, the *flawless detection* criterion (Equation (6)) could suppress the estimation error for mean heart rate (**μ**). This was not the case, however, for the HRV features (**σ**, **rmssd**, **lf** and **hf**), which still showed non-zero error rates even when no sample was missing. As stated earlier in this study, these residual errors may arise from two

distinct phenomena: (i) the unavoidable noise in IBI estimation due to the smooth shape of the PPG waveform, or (ii) the varying pulse transit between heart and the wrist. Since factor (i) varies from one heartbeat to another, it may explain the residual error rates found for short-term HRV features (**rmssd**, **hf**). Since factor (ii) depends on physiological variables like blood pressure, it may explain the residual error rates found for longer-term HRV features (**σ, lf**). Another striking result is that those residual errors are typically lower for longer-term (**σ, lf**) than for shorter-term (**rmssd**, **hf**) HRV features. Hence, Whether or not pulse rate variability and HRV should be considered as distinct cardiac measures is still an open question. The answer depends on which features have to be computed and which error rates can be expected without compromising higher-level computations.

As illustrated in this study, there is still work to be done on cardiac monitoring systems (using either PPG or ECG) to make the estimation of heart rate estimation more robust to real-life conditions. Cardiac measurement will likely remain prone to artefacts, in spite of the technical advances made in the recent years. This illustrates the need for system-specialized SQIs (like the one proposed by BioHarness) to ensure that higher-level algorithms – namely, inner states monitoring or heart event prevention – will not run on false information.

In advanced cardiac monitoring applications, however, there is also need for white-box indicators to implement strong quality management strategies in the feature extraction process. In that perspective, the current study proposed a method to timely control the risk of error due to data loss in real-life settings. At a time when cardiac monitoring techniques are being developed beyond ECG and PPG, (remote PPG for instance, which is useful when body contact is not desired), this method should help the researcher to characterize several sensors outputs comparatively.

5. Patents

The work reported in this manuscript has led to patent applications, currently registered under N° FR3098390 - EP3763283 - US2021007674.

Author Contributions: Conceptualization, C.G, S.C and A.C.; methodology, G.V.; software, G.V.; investigation, G.V.; data curation, G.V.; writing—original draft preparation, G.V.; writing—review and editing, C.G., S.C., A.C. and G.V.; supervision, S.C., A.C. and C.G; funding acquisition, C.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the the Ethics Committee in Non-Interventional Research (CERNI) related to COMUE Univ. Grenoble-Alpes, under agreement N° 2015-05-12-67.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data reported in this study is not currently available in a public database.

Conflicts of Interest: The authors declare no conflict of interest in citing any device or research material in the current manuscript.

References

1. Swan, M. The Quantified Self: Fundamental Disruption in Big Data Science and Biological Discovery. *Big Data* **2013**, *1*, 85–99, doi:10.1089/big.2012.0002.
2. Picard, R.W. *Affective Computing*; Citeseer, 1997;
3. Greene, S.; Thapliyal, H.; Caban-Holt, A. A Survey of Affective Computing for Stress Detection: Evaluating Technologies in Stress Detection for Better Health. *IEEE Consum. Electron. Mag.* **2016**, *5*, 44–56, doi:10.1109/MCE.2016.2590178.
4. Schmitt, P.M.; Gehin, C.; Delhomme, G.; McAdams, E.; Dittmar, A. Flexible Technologies and Smart Clothing for Citizen Medicine, Home Healthcare, and Disease Prevention. *IEEE Trans. Inf. Technol. Biomed.* **2005**, *9*, 325–336, doi:10.1109/TITB.2005.854505.

5. Yilmaz, T.; Foster, R.; Hao, Y. Detecting Vital Signs with Wearable Wireless Sensors. *Sensors* **2010**, *10*, 10837–10862, doi:10.3390/s101210837.
6. Acharya, U.R.; Joseph, K.P.; Kannathal, N.; Lim, C.M.; Suri, J.S. Heart Rate Variability: A Review. *Med. Biol. Eng. Comput.* **2006**, *44*, 1031–1051, doi:10.1007/s11517-006-0119-0.
7. Kim, H.-G.; Cheon, E.-J.; Bai, D.-S.; Lee, Y.H.; Koo, B.-H. Stress and Heart Rate Variability: A Meta-Analysis and Review of the Literature. *Psychiatry Investig.* **2018**, *15*, 235–245, doi:10.30773/pi.2017.08.17.
8. Fairclough, S.H.; Mulder, L.J.M. Psychophysiological Processes of Mental Effort Investment. *Motiv. Affects Cardiovasc. Response Mech. Appl.* **2011**, 61–76.
9. Vandecasteele, K.; De Cooman, T.; Gu, Y.; Cleeren, E.; Claes, K.; Paesschen, W.V.; Huffel, S.V.; Hunyadi, B. Automated Epileptic Seizure Detection Based on Wearable ECG and PPG in a Hospital Environment. *Sensors* **2017**, *17*, 2338, doi:10.3390/s17102338.
10. Gruetzmann, A.; Hansen, S.; Müller, J. Novel Dry Electrodes for ECG Monitoring. *Physiol. Meas.* **2007**, *28*, 1375–1390, doi:10.1088/0967-3334/28/11/005.
11. Fuhrhop, S.; Lamparth, S.; Kirst, M.; Wagner, G. v.; Ottenbacher, J. Ambulant ECG Recording with Wet and Dry Electrodes: A Direct Comparison of Two Systems. In Proceedings of the World Congress on Medical Physics and Biomedical Engineering, September 7 - 12, 2009, Munich, Germany; Dössel, O., Schlegel, W.C., Eds.; Springer Berlin Heidelberg, 2009; pp. 305–307.
12. Sinex, J.E. Pulse Oximetry: Principles and Limitations. *Am. J. Emerg. Med.* **1999**, *17*, 59–66, doi:10.1016/S0735-6757(99)90019-0.
13. Gil, E.; Orini, M.; Bailón, R.; Vergara, J.M.; Mainardi, L.; Laguna, P. Photoplethysmography Pulse Rate Variability as a Surrogate Measurement of Heart Rate Variability during Non-Stationary Conditions. *Physiol. Meas.* **2010**, *31*, 1271–1290, doi:10.1088/0967-3334/31/9/015.
14. Schäfer, A.; Vagedes, J. How Accurate Is Pulse Rate Variability as an Estimate of Heart Rate Variability?: A Review on Studies Comparing Photoplethysmographic Technology with an Electrocardiogram. *Int. J. Cardiol.* **2013**, *166*, 15–29, doi:10.1016/j.ijcard.2012.03.119.
15. Thomas, S.S.; Nathan, V.; Zong, C.; Soundarapandian, K.; Shi, X.; Jafari, R. BioWatch: A Noninvasive Wrist-Based Blood Pressure Monitor That Incorporates Training Techniques for Posture and Subject Variability. *IEEE J. Biomed. Health Inform.* **2016**, *20*, 1291–1300, doi:10.1109/JBHI.2015.2458779.
16. Periyasamy, V.; Pramanik, M.; Ghosh, P.K. Review on Heart-Rate Estimation from Photoplethysmography and Accelerometer Signals During Physical Exercise. *J. Indian Inst. Sci.* **2017**, *97*, 313–324, doi:10.1007/s41745-017-0037-1.
17. Sweeney, K.T.; Kearney, D.; Ward, T.E.; Coyle, S.; Diamond, D. Employing Ensemble Empirical Mode Decomposition for Artifact Removal: Extracting Accurate Respiration Rates from ECG Data during Ambulatory Activity. In Proceedings of the 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); July 2013; pp. 977–980.
18. del Río, B.A.S.; Lopetegi, T.; Romero, I. Assessment of Different Methods to Estimate Electrocardiogram Signal Quality. In Proceedings of the 2011 Computing in Cardiology; IEEE, 2011; pp. 609–612.
19. Behar, J.; Oster, J.; Li, Q.; Clifford, G.D. A Single Channel ECG Quality Metric. In Proceedings of the 2012 Computing in Cardiology; September 2012; pp. 381–384.
20. Li, Q.; Clifford, G.D. Dynamic Time Warping and Machine Learning for Signal Quality Assessment of Pulsatile Signals. *Physiol. Meas.* **2012**, *33*, 1491–1501, doi:10.1088/0967-3334/33/9/1491.
21. Papini, G.B.; Fonseca, P.; Aubert, X.L.; Overeem, S.; Bergmans, J.W.M.; Vullings, R. Photoplethysmography Beat Detection and Pulse Morphology Quality Assessment for Signal Reliability Estimation. In Proceedings of the 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); July 2017; pp. 117–120.
22. Orphanidou, C.; Drobnjak, I. Quality Assessment of Ambulatory ECG Using Wavelet Entropy of the HRV Signal. *IEEE J. Biomed. Health Inform.* **2017**, *21*, 1216–1223, doi:10.1109/JBHI.2016.2615316.
23. Orphanidou, C.; Bonnici, T.; Charlton, P.; Clifton, D.; Vallance, D.; Tarassenko, L. Signal-Quality Indices for the Electrocardiogram and Photoplethysmogram: Derivation and Applications to Wireless Monitoring. *IEEE J. Biomed. Health Inform.* **2015**, *19*, 832–838, doi:10.1109/JBHI.2014.2338351.
24. How Is IBI.Csv Obtained? Available online: http://support.empatica.com/hc/en-us/articles/201912319-How-is-IBI-csv-obtained- (accessed on 1 April 2019).
25. Ollander, S.; Godin, C.; Campagne, A.; Charbonnier, S. A Comparison of Wearable and Stationary Sensors for Stress Detection. In Proceedings of the Systems, Man, and Cybernetics (SMC), 2016 IEEE International Conference on; IEEE, 2016; pp. 004362–004366.
26. Schuurmans, A.A.T.; de Looff, P.; Nijhof, K.S.; Rosada, C.; Scholte, R.H.J.; Popma, A.; Otten, R. Validity of the Empatica E4 Wristband to Measure Heart Rate Variability (HRV) Parameters: A Comparison to Electrocardiography (ECG). *J. Med. Syst.* **2020**, *44*, 190, doi:10.1007/s10916-020-01648-w.
27. Milstein, N.; Gordon, I. 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.* **2020**, *14*, 148, doi:10.3389/fnbeh.2020.00148.
28. Johnstone, J.A.; Ford, P.A.; Hughes, G.; Watson, T.; Mitchell, A.C.S.; Garrett, A.T. Field Based Reliability and Validity of the Bioharness™ Multivariable Monitoring Device. *J. Sports Sci. Med.* **2012**, *11*, 643–652.
29. Kahneman, D.; Krueger, A.B.; Schkade, D.A.; Schwarz, N.; Stone, A.A. A Survey Method for Characterizing Daily Life Experience: The Day Reconstruction Method. *Science* **2004**, *306*, 1776–1780, doi:10.1126/science.1103572.

30. Pan, J.; Tompkins, W.J. A Real-Time QRS Detection Algorithm. *IEEE Trans. Biomed. Eng.* **1985**, *BME-32*, 230–236, doi:10.1109/TBME.1985.325532.
31. Sedghamiz, H. *Matlab Implementation of Pan Tompkins ECG QRS Detector.*; 2014;
32. Kemper, K.J.; Hamilton, C.; Atkinson, M. Heart Rate Variability: Impact of Differences in Outlier Identification and Management Strategies on Common Measures in Three Clinical Populations. *Pediatr. Res.* **2007**, *62*, 337.
33. Karlsson, M.; Hörnsten, R.; Rydberg, A.; Wiklund, U. Automatic Filtering of Outliers in RR Intervals before Analysis of Heart Rate Variability in Holter Recordings: A Comparison with Carefully Edited Data. *Biomed. Eng. Online* **2012**, *11*, 2.
34. Lomb, N.R. Least-Squares Frequency Analysis of Unequally Spaced Data. *Astrophys. Space Sci.* **1976**, *39*, 447–462, doi:10.1007/BF00648343.
35. VanderPlas, J.T. Understanding the Lomb–Scargle Periodogram. *Astrophys. J. Suppl. Ser.* **2018**, *236*, 16, doi:10.3847/1538-4365/aab766.
36. Bouten, C.V.C.; Koekkoek, K.T.M.; Verduin, M.; Kodde, R.; Janssen, J.D. A Triaxial Accelerometer and Portable Data Processing Unit for the Assessment of Daily Physical Activity. *IEEE Trans. Biomed. Eng.* **1997**, *44*, 136–147, doi:10.1109/10.554760.
37. Karantonis, D.M.; Narayanan, M.R.; Mathie, M.; Lovell, N.H.; Celler, B.G. Implementation of a Real-Time Human Movement Classifier Using a Triaxial Accelerometer for Ambulatory Monitoring. *IEEE Trans. Inf. Technol. Biomed.* **2006**, *10*, 156–167.
38. Shaffer, F.; Ginsberg, J.P. An Overview of Heart Rate Variability Metrics and Norms. *Front. Public Health* **2017**, *5*, 258, doi:10.3389/fpubh.2017.00258.