TAMPEREEN TEKNILLINEN YLIOPISTO
TAMPERE UNIVERSITY OF TECHNOLOGY

Jakub Parak

Evaluation of Wearable Optical Heart Rate Monitoring Sensors

Julkaisu 1580 • Publication 1580

Tampere 2018

Tampereen teknillinen yliopisto. Julkaisu 1580
Tampere University of Technology. Publication 1580

Jakub Parak

Evaluation of Wearable Optical Heart Rate Monitoring Sensors

Thesis for the degree of Doctor of Science in Technology to be presented with due permission for public examination and criticism in Rakennustalo Building, Auditorium RN201, at Tampere University of Technology, on the 8th of November 2018, at 12 noon.

Tampereen teknillinen yliopisto - Tampere University of Technology
Tampere 2018

Doctoral candidate:	Jakub Parak Personal Health Informatics Research Group Faculty of Biomedical Sciences and Engineering Tampere University of Technology Finland
Supervisor:	Adjunct Professor Ilkka Korhonen Personal Health Informatics Research Group Faculty of Biomedical Sciences and Engineering Tampere University of Technology Finland
Instructor:	Professor Pavel Sovka Department of Circuit Theory Faculty of Electrical Engineering Czech Technical University in Prague Czech Republic
Pre-examiners:	Docent Mika Tarvainen Department of Applied Physics University of Eastern Finland Finland
	Docent Ari Nummela Research Institute for Olympic Sports Finland
Opponent:	Professor Oliver Amft Chair of Digital Health Friedrich-Alexander Universität Erlangen-Nürnberg Germany

ISBN 978-952-15-4210-7 (printed)
ISBN 978-952-15-4246-6 (PDF)
ISSN 1459-2045

Abstract

Heart rate monitoring provides valuable information about an individual's physiological condition. The information obtained from heart rate monitoring can be used for a wide range of purposes such as clinical diagnostics, assessment of the efficiency of training for sports and fitness, or of sleep quality and stress levels in wellbeing applications. Other useful parameters for describing a person's fitness, such as maximal oxygen uptake and energy expenditure, can also be estimated using heart rate measurement. The traditional 'gold standard' for heart rate monitoring is the electrocardiograph, but nowadays there are a number of alternative methods too. Of these, optical sensors provide a relatively simple, low-cost and unobtrusive technology for monitoring heart rate and they are widely accepted by users. There are many factors affecting the measurement of optical signals that have an effect on the accuracy of heart rate estimation. However, there is a lack of standardized and unified methodology for comparing the accuracy of optical heart rate sensors to the 'gold standard' methods of measuring heart rate. The widespread use of optical sensors for different purposes has led to a pressing need for a common objective methodology for the evaluation of how accurate these sensors are. This thesis presents a methodology for the objective evaluation of optical heart-rate sensors. The methodology is applied in evaluation studies of four commercially available optical sensors. These evaluations were carried out during both controlled and non-controlled sporting and daily life activities. In addition, evaluation of beat detection accuracy was carried out in non-controlled sleep conditions. The accuracy of wrist-worn optical heart-rate sensors in estimating of maximal oxygen uptake during submaximal exercise and energy expenditure during maximal exercise using heart rate as input parameter were also evaluated. The accuracy of a semi-continuous heart rate estimation algorithm designed to reduce power consumption for long-term monitoring was also evaluated in various conditions. The main findings show that optical heart-rate sensors may be highly accurate during rhythmic sports activities, such as jogging, running, and cycling, including ramp-up running during maximal exercise testing. During non-rhythmic activities, such as intermittent hand movements, the sensors' accuracy depends on where they are worn. During sleep and motionless conditions, the optical heart-rate sensors' estimates for beat detection and inter-beat interval showed less than

one percent inaccuracy against the values obtained using standard measurement techniques. The sensors were also sufficiently accurate at measuring the inter-beat intervals to be used for calculating the heart rate variability parameters. The estimation accuracy of the fitness parameters derived from measured heart rate can be described as follows. An assessment of the maximal oxygen uptake estimation during a sub-maximal outdoor exercise had a precision close to a sport laboratory measurement. The energy expenditure estimation during a maximal exercise was more accurate during higher intensity of exercise above aerobic threshold but the accuracy decreased at lower intensity of exercise below the aerobic threshold, in comparison with the standardized reference measurement. The semi-continuous algorithm was nearly as accurate as continuous heart-rate detection, and there was a significant reduction in the power consumption of the optical chain components up to eighty percent. The results obtained from these studies show that, under certain conditions, optical sensors may be similarly accurate in measuring heart rate as the ‘gold standard’ methods and they can be relied on to monitor heart rate for various purposes during sport, everyday activities, or sleep.

"*We choose to go to the Moon! … We choose to go to the Moon in this decade and do the other things, not because they are easy, but because they are hard, because that goal will serve to organize and measure the best of our energies and skills, because that challenge is one that we are willing to accept, one we are unwilling to postpone, and one we intend to win..."*

John Fitzgerald Kennedy speech in Houston, on September 12, 1962

Preface

The research presented in this thesis was carried out at between 2013 and 2017 at Tampere University of Technology (TUT) in cooperation with PulseOn, Oy. Throughout this time I have had the pleasure of working in a scientific environment full of enthusiastic and talented people who have enriched my research with their skills, experience and ideas.

First, I would like to express my gratitude to my supervisor, adjunct professor Dr. Ilkka Korhonen, for initially introducing me to this research field, and for his great ideas and continuous support during all of my research activities and doctoral studies. He has inspired me with his help in developing the basic concept for my research and has continuously encouraged me in the writing of it. I would also like to express my gratitude to my instructor, professor Dr. Pavel Sovka, for his primary introduction into the research environment and his lucid explanations of the main scientific principles, approaches and methods.

I also wish to thank my pre-examiners docent Dr. Ari Nummela and docent Dr. Mika Tarvainen for their valuable comments and objective criticism of my thesis.

My gratitude also goes to Dr. Adrian Tarniceriu from PulseOn, the co-author of most of my publications, who has taught me how to express my scientific research in a suitable written form for publication. Indeed, several of the ideas presented in this thesis are the direct outcome of our intense discussions, and I would also like to thank him for his invaluable comments on the final version of this thesis. I am also grateful to Dr. Ricard Delgado-Gonzalo and Dr. Philippe Renevey, both from the Swiss Center for Electronics and Microtechnology, for co-authoring some of my publications.

My thanks also go to all the members of the Personal Health Informatics research group at TUT, especially to Julia Pietilä for her encouragement during the final write-up of this thesis, and to Dr. Hannu Nieminen for maintaining and coordinating the practical organization of the research group. I would also like to thank the research assistants, Aleksi Haavikko and Maria Uuskoski for their assistance with the data collection campaigns and Jan Machek for his cooperation with the data processing. I would like to express my gratitude to all my colleagues in PulseOn for providing general support and material resources, especially to Marko Nurmi for his helpful and friendly collaboration and for bringing his practical engineering experience to this research area. My thanks also go to all of the volunteers who participated as test subjects in the evaluation studies, and I am grateful to Varala Sports Center in Tampere and the VTT Technical Research Centre of

Finland in Tampere for providing me with free laboratory premises in which to carry out my experiments.

My thanks also go to all my colleagues and friends: Pavel Marek and Jakub Esner for their good company and for always being willing test subjects for my research experiments, Martin Horak and Andrej Zitnan for providing long-distance support for their empathy and their intellectual support, Jakub Kocmanek for his stylistic advice during the writing of this thesis, and Jan Sedlak for our lengthy discussions of the practical aspects of scientific research. I would also like to thank Kaisa and Johan Plomp for providing me with a stable and supportive home and family environment during my doctoral studies here in Tampere. Finally, I want to express my gratitude to my family, especially to my mother and sister for their constant encouragement which has led to the successful completion of this doctoral thesis.

This work has been financially supported by the Finnish Funding Agency for Innovation (TEKES) for Wellness Ecosystem project executed in 2013 and by the doctoral student grant provided by Tampere University of Technology in 2016 and 2017.

Tampere, August 2018

Jakub Parak

Contents

List of Abbreviations

AC – alternate current

AFE – analogue frontend

ANS – autonomic nervous system

BA – Bland-Altman

BMI – body mass index

BMR – basal metabolic rate

DC – direct current

ECG – electrocardiography

EE – energy expenditure

GPS – global positioning system

HR – heart rate

HR_{max} – maximum heart rate

HRV – heat rate variability

IBI – inter-beat interval

IC – indirect calorimetry

ISO – International Organization for Standardization

LED – light emitting diode

LoA – limits of agreement

MAE – mean absolute error

MAP – mean arterial pressure

MAPE – mean absolute percentage error

ME – mean error

MPE – mean percentage error

OHR – optical heart rate

PD – photodetector

PPG – photoplethysmography

RMSE – root mean square error

RMSE(D) – root mean squared error (difference)

RMSSD – root mean square of the successive differences

RRI – RR interval

SD – standard deviation

SE – standard error

SEE – standard error of estimate

SEM – standard error of mean

SPO_2 – blood oxygen saturation

TEE – total energy expenditure

VO_2 – oxygen uptake

VO_{2max} – maximal oxygen uptake

List of Publications

I. Parak, J. & Korhonen, I. (2014). Evaluation of wearable consumer heart rate monitors based on photopletysmography, 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 3670-3673.

II. Delgado-Gonzalo, R., Parak, J., Tarniceriu, A., Renevey, P., Bertschi, M. & Korhonen, I. (2015). Evaluation of accuracy and reliability of PulseOn optical heart rate monitoring device, 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 430-433.

III. Parak, J., Tarniceriu, A., Renevey, P., Bertschi, M., Delgado-Gonzalo, R. & Korhonen, I. (2015). Evaluation of the beat-to-beat detection accuracy of PulseOn wearable optical heart rate monitor, 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 8099-8102.

IV. Tarniceriu, A., Parak, J., Renevey, P., Nurmi, M., Bertschi, M., Delgado-Gonzalo, R. & Korhonen, I. (2016). Towards 24/7 Continuous Heart Rate Monitoring, 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 186-189.

V. Parak, J., Uuskoski, M., Machek, J. & Korhonen, I. (2017). Estimating Heart Rate, Energy Expenditure, and Physical Performance With a Wrist Photoplethysmographic Device During Running, JMIR mHealth and uHealth, 5(7), e97.

Author’s contribution

I. The author had the main responsibility for designing the evaluation campaign and for carrying out the experiments. He had the main responsibility for data processing and calculation of the results, and was the main author of this publication.

II. The author had primary responsibility for updating the evaluation protocol, supervising data collection in the laboratory and for performing and supervising the outdoor experiments. The calculation and interpretation of the results was performed with the help of A. Tarniceriu and R. Delgado-Gonzalo, who co-authored this publication.

III. The author had primary responsibility for the data collection, including instructing the participants on how to perform non-controlled measurements. The author shared responsibility for the design of the evaluation methodology and the calculation of the results with A. Tarniceriu and was the main author of this publication.

IV. The author had primary responsibility for preparing the test datasets used for the evaluation of a semi-continuous heart rate estimation algorithm. He shared responsibility for the design and evaluation of the algorithm with M. Nurmi and A. Tarniceriu, who co-authored this publication.

V. The author shared responsibility for designing and executing the evaluation campaign with M. Uuskoski. He shared responsibility for the data analyses, and for the processing and estimation of the final results with J. Machek and M. Uuskoski. The author is the main author of this publication.

1 Introduction

Heart rate (HR) is one of the most fundamental techniques for measuring vital signs and has been used almost since the dawn of civilization. For example, in ancient Greece, the physician and scientist Herophilos (ca. 335 to ca. 280 BC) measured HR by timing the pulse using a portable clepsydra (a water clock) (Bedford 1951; Bay & Bay 2010). In the 17th century, English physician John Floyer constructed "The Physician Pulse Watch" and introduced quantitative HR measurement by counting the number of beats per minute (Floyer 1707; Floyer 1710). The modern era of HR monitoring was initiated in the late 19th century by William Einthoven's invention of a technique for measuring the electrical activity of the heart (Einthoven 1895). By the late 1930s, Alrick Hertzman performed the first experimental measurements of blood flow with a photoelectric plethysmograph (Hertzman 1937; Hertzman 1938). The real milestone in the history of wearable HR monitoring, however, came in 1960, when Norman Holter constructed the first portable electrocardiographic recorder (Holter 1961). HR monitoring is widely used nowadays in various activities, such as sport, but also for assessing people's general fitness and wellbeing, i.e. healthcare. In professional sport training, especially in endurance sports, HR measurement is a common support tool for tracking an athlete's physical condition in order to design effective, intensive training programmes which detect and prevent overexertion and include suitable recovery periods (Achten & Jeukendrup 2003). It is difficult to measure maximal oxygen uptake (VO_{2max}) and energy expenditure (EE) in field exercise conditions using standard measurement techniques, but the values can be estimated from their relationship with HR (Achten & Jeukendrup 2003).

For much the same reasons, HR monitoring can be extremely beneficial in monitoring the average person's recreational sport or daily fitness activities. As they are based on the physiological response of the whole body, analyses of heart rate variability (HRV) derived from beat-by-beat HR monitoring can provide valuable information about a person's general fitness and current level of physical activity (Mutikainen et al. 2014; Hallman et al. 2015), psychological stress (Teisala et al. 2014; Kaikkonen et al. 2017; Fohr

et al. 2017), and sleep quality and recovery (Myllymaki et al. 2012; Tobaldini et al. 2013; Pietilä et al. 2015). In health care, besides its fundamental diagnostic use for monitoring a patient's vital signs, HR monitoring can also be used for other purposes. It is ideal for monitoring the health of out-patients in remote home monitoring, thus enabling early intervention (Merilahti et al. 2009), or as a predictor for various cardiac diseases (Agewall et al. 2017). Additionally, when combined with further HRV analysis, it is a useful tool for examining the autonomic nervous system (ANS) (Malik 1996; Sztajzel 2004).

Nowadays, most of the HR monitoring during sport and daily activities is usually carried out with devices based on the electrocardiography (ECG) principle. These involve the use of chest straps and disposable electrode recorders, and most of them are indeed highly accurate. The key benefits of ECG-based HR recorders are their straightforward electrical signal acquisition from the body, their relatively simple digital processing and their robustness to the effects of motion, especially with regard to the way the devices are constructed. Although ECG devices are relatively simple to manufacture, they do restrict the wearers. The contact between the skin and the sensors needs to be sufficiently moist to provide good conductivity, which is particularly apparent before starting exercise or during long-term monitoring. The wearer's skin might be irritated by the adhesive used to attach the disposable electrodes, or by the material used for the chest strap, particularly during long-term measurements, because a relatively large area of skin has to be covered by these sensors. In addition, the position of the chest straps can be obtrusive, especially for females.

The key benefits of optical-based wearable HR monitors are their small size and their common and familiar wearing position on the wrist. However, optical HR (OHR) monitors have their own limitations, in that the quality of the signal can be affected by the wearer's skin color, the ambient light, and the type and degree of motion.

Despite their limitations, OHR monitors are becoming more and more popular with the general public. In addition, they are being utilised more for research into medical health and fitness because of their relative ease of use. However, there are no standardized guidelines or procedures on how to objectively evaluate these devices, nor experimental platforms for evaluating new medical products. Bassett et al. (Bassett et al. 2012) have already emphasized the necessity of testing the fundamental accuracy of wearable sensors used for scientific purposes, and considering their increasingly widespread use, the importance of such an objective is growing.

The main objective of this thesis is to develop an objective methodology for assessing the accuracy of wearable OHR sensors, and to apply this methodology to evaluate the accuracy of selected commercially available high-end consumer OHR sensors.

This thesis is based on five recent research papers published between 2013 and 2017 which focused on measuring the accuracy of OHR sensors. Three of the publications focus on evaluating commercially available OHR monitors during sport activities. In addition, one of the publications presents the beat-to-beat interval detection accuracy of OHR, another shows how power consumption in OHR can be reduced and one describes estimation accuracy of the EE and VO_{2max} derived from OHR measurements.

Following this introductory chapter, Chapter 2 presents the objectives of this thesis. Chapter 3 describes the basic physiology of the heart and HR. It then describes the principles on which OHR is based, and the principles used for estimating VO_{2max} and EE based on HR. Chapter 4 summarizes and analyzes the methodologies used in previous studies of optical wearable sensors. Chapter 5 describes an 'evaluation framework' for the objective evaluation of OHR. Chapter 6 summarizes the results of the cited publications addressing specific objectives. Chapter 7 discusses the results, the impact and limitations of the studies, and possible future directions for further research, and Chapter 8 summarises the general findings and conclusions.

2 Objectives of the thesis

The main objectives of this thesis are to develop an objective evaluation methodology for assessing the accuracy of wearable OHR sensors, and to apply this methodology to the evaluation of the accuracy of selected, commercially-available, high-end, consumer OHR sensors. The specific objectives of the thesis are:

1. To develop an objective OHR evaluation methodology that can be used for the evaluation of OHR accuracy in various real life situations (Publications I-V)
2. To evaluate the accuracy of selected high-end OHR devices during sports (Publications I, II and V)
3. To evaluate the beat-to-beat accuracy of a selected OHR device during sleep (Publication III)
4. To evaluate the accuracy of EE and VO_{2max} estimation based on OHR and mobile phone-based speed estimation (Publication V)
5. To evaluate a low-power approach for OHR estimation during everyday use (Publication IV)

3 Physiology and measurement principles

3.1 Heart and heart rate

The heart is a muscular body organ that pumps blood circulating through vessels in the body (Tortora & Grabowski 2003). It is located in the mediastinum and consists of four chambers; the right atrium and ventricle, and the left atrium and ventricle (Tortora & Grabowski 2003). Blood circulation is important for transporting various substances around the body, such as oxygen, carbon dioxide and nutrients, and is vital for regulating life processes (Tortora & Grabowski 2003).

HR is defined as the number of heart contractions per unit of time. The contractions are induced by pacemaker cells in the sinoatrial (SA) node of the heart which the other cardiac muscle cells follow (Vander et al. 1990). The inherent SA heart rhythm rate in the absence of any neural or hormonal influences is 100 beats per minute (bpm). HR is typically regulated by sympathetic (stimulating) and parasympathetic (inhibiting) activity of the ANS. The normal resting HR is below 100 bpm because parasympathetic activity has more influence during rest (Vander et al. 1990). In addition, heart rate activity may also be modified by various chemicals, drugs, hormones and ions (Marieb 2006).

Each contraction of the left ventricle ejects blood into the aorta. This results in increased blood pressure caused by the hydrostatic pressure exerted by the blood against the inner walls of the vessels (Tortora & Grabowski 2003). The factors affecting the magnitude of the pulse (the difference between systolic and diastolic pressure) are stroke volume, speed of stroke, ejection volume and arterial compliance (Vander et al. 1990). Stroke volume depends on the volume of blood in the ventricles before contraction, and the input amplitude of the sympathetic nervous system for ventricle contractions (Vander et al. 1990). The blood pressure wave is propagated through the arterial system to the peripheral arteries (arterioles and capillaries). The alternating expansion and recoil of the

elastic arteries creates a pressure wave, which is the pulse (Tortora & Grabowski 2003). In a healthy person, the pulse rate (pressure surges per minute) equals the HR (beats per minute) (Marieb 2006). There are certain pathophysiological states, such as heart arrhythmias (e.g. ventricular fibrillation or sudden heart arrest) or arterial embolism (e.g. thromboembolism), which can result in the absence of a pulse in the arteries (Tortora & Grabowski 2003). The relative distribution of blood flow to the particular organs is regulated by the change of resistance in the arterioles. In the cardiovascular system, this resistance is a measure of the friction between the blood and the walls of the vessels, which can impede the flow of the blood (Vander et al. 1990). Blood flow rate to a particular organ F_{organ} is directly proportional to the mean arterial pressure (*MAP*) and the resistance of the organ R_{organ} (1).

$$F_{organ} = \frac{MAP}{R_{organ}} \tag{1}$$

Large arteries are reservoirs of pressure in the body (Vander et al. 1990). The resistance change in the arterioles is based on the adjustment of their diameter using a smooth muscle. When the muscle relaxes it increases the diameter (vasodilation) and when it contracts it decreases the diameter of the vessel (vasoconstriction) (Vander et al. 1990). These changes are controlled by both local and extrinsic control mechanisms. Local control mechanisms regulate the blood flow under the following conditions: increased metabolic activity (active hyperemia), which usually demands an increase of blood flow in organ tissue; pressure changes, which require the maintenance of a constant blood flow (pressure autoregulation); an increase of blood flow after blood supply occlusion (reactive hyperemia); or, vasodilatation, which occurs during inflammation in response to an injury (Vander et al. 1990). The extrinsic control mechanisms are based on sympathetic nerves controlling blood flow in the skin. In cold weather for example, they trigger a reflex increase in sympathetic activity which causes vasoconstriction and reduced blood flow to the skin, making the skin feel cold to the touch (Vander et al. 1990; Marieb 2006) This reduction in blood flow, called thermoregulation, is the body's way of retaining the heat in the warm blood for the body's internal organs (Tortora & Grabowski 2003). In contrast, in warm weather the increased body temperature inhibits the sympathetic activity, causing the arterioles to dilate, which enhances circulation in the skin (Vander et al. 1990; Marieb 2006). The regulation of blood flow due to vasoconstriction and vasodilation have a significant impact on the signal quality in photoplethymographic blood flow measurements (Kamal et al. 1989).

A normal HR, or pulse rate, is about 75 beats per minute (bpm) for a body at rest. Rapid HR, over 100 bpm, is called tachycardia. Bradycardia is the opposite, meaning a slower than normal HR, typically less than 60 bpm (Marieb 2006) and can often be observed in endurance athletes (Tortora & Grabowski 2003). The main physical factors that have an influence on HR include gender, age, exercise, stress, body temperature or physical condition. HR is usually faster in females than males (72 – 80 bpm vs 64 – 72 bpm, respectively), while the resting HR of the human foetus is around 140 – 160 bpm. Heat can boost the metabolic rate of the heart cells, while cold has an opposite effect and decreases the HR. Physical exercise or stress can temporarily stimulate nervous controls (sympathetic control) and this can also increase the HR (Marieb 2006).

Electrocardiography (ECG or EKG) is the standard method for detecting heart activity. The principle behind it is to measure the electrical potential generated by the heart muscles during contraction. An electrocardiogram is the recorded signal of heart activity consisting of 5 waves: P, Q, R, S, and T (Figure 1) (Tortora & Grabowski 2003). Inter-beat interval (IBI) or RR interval (RRI) is the time between two successive beats (R waves) and is measured in milliseconds. Pulse time, derived from the pulse wave, also corresponds to RRI (Figure 2) (Lemay et al. 2014). Beat detection in an ECG signal for RR interval estimation can usually be realized by detecting the QRS complex (Pan & Tompkins 1985). The RR interval and the HR value have a non-linear relationship (1/RRI) (Korhonen 1997). HRV is the natural variation in time between consecutive beats, and is predominantly determined by the extrinsic regulation of the ANS (Shaffer & Ginsberg 2017). HRV analyses in the time and frequency domain can be utilized for various applications, such as clinical practice (Malik 1996; Sztajzel 2004), sleep quality measurement (Myllymaki et al. 2012), and sport (Aubert et al. 2003).

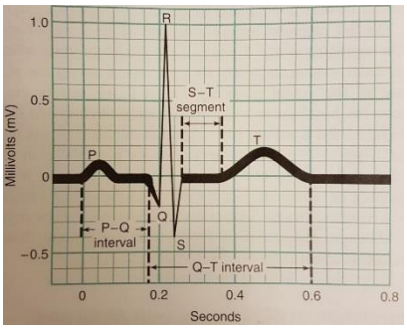


Figure 1: ECG signal waveform
From (Tortora & Grabowski 2003). © 2003 by Biological Sciences Textbooks, Inc. and Sandra Reynolds Grabowski. Reprinted by permission of John Wiley & Sons, Inc.

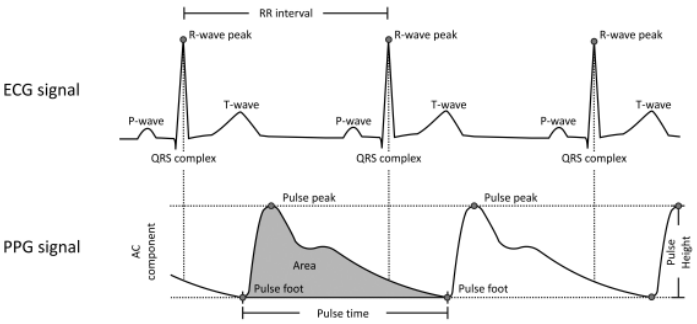


Figure 2: ECG and PPG signal, RR interval and pulse time. Reprinted from (Lemay et al. 2014) © 2014 with permission from Elsevier.

3.2 PPG measurement principle

Photoplethysmography (PPG) is a simple and low-cost, non-invasive, optical measurement technology. The principle behind it is to measure the light propagation in tissue during the cardiac cycle by detecting the volume of blood in the microvascular bed tissue, which changes with the blood flow (Challoner & Ramsay 1974; Allen 2007). PPG technology is based on a light beam illuminating the tissue. Some of the light is absorbed in the tissue, while some of it is reflected or trans-illuminated to the optical sensors (Lemay et al. 2014). PPG has complicated relationships to several biomechanical, optical and physiological covariates (Allen 2007; Reisner et al. 2008) but can provide useful information about cardiovascular and ANS activity (Kamal et al. 1989).

The Beer-Lamber law (2) describes the attenuation of transmitted light from the source of the light beam to the optical sensor. When a monochromatic light beam, $I_{0,}$ propagates in a homogeneous medium, the light intensity, I, decreases exponentially as a function of the path length, l, and the light absorption coefficient, α, which is related to the medium's properties at a specific wavelength.

$$I = I_0 e^{\alpha l} \tag{2}$$

The Beer-Lambert law applies to multiple substances absorbing light in a medium, or for a sequence of several different media. In both cases, the total absorbance is expressed as the sum of the absorbencies of the individual components. According to the Beer-Lambert law, the sum of the transmitted and absorbed light is equal to the incident source light, thus not taking into account scattering or reflection of light. Hence, it is a simplification of the actual physical process.

Because the fundamental Beer-Lambert law expresses the absorbance of light propagated through homogeneous layers (Reisner et al. 2008; Lemay et al. 2014), it cannot be directly applied to the absorbance of light in biological structures such as blood, skin and other biological tissues, as they are inhomogeneous. This inhomogeneity leads to the non-linear absorbance of light and causes complex changes in the light's reflection and absorption, mainly due to movement or to variations in the inhomogeneous structures (Lemay et al. 2014). The absorbance and scattering of light is also subject to the orientation of the red blood cells, which depends on the cardiac cycle (Nijboer et al. 1981; Lemay et al. 2014). Living tissue with a blood flow can be modeled as a concatenation

of several media (skin layers, tissue, blood, arteries, veins, inter-cellular liquids etc.) or blood components (oxyhemoglobin, deoxyhemoglobin) that are characterized by particular path lengths and light absorption coefficients (Bronzino 1995; Lemay et al. 2014).

In an approximate model, one layer of a medium represents the veins and arteries, which change the absorption of light and the attenuation of transmitted light with pulsatile blood propagation invoked by the heartbeat. Increases of the pulsatile blood pressure in the vessels modifies their geometry (due to volume change) and their optical properties (due to changes in blood composition and concentration) (Lemay et al. 2014). It is the volumetric changes of the venous and arterial blood which are the origin of PPG signal variations (Figure 3) (Lemay et al. 2014). They are usually divided into AC and DC signal components. In PPG signals, pulsatile arterial blood is represented with the AC component (Challoner & Ramsay 1974), while the “constant” light absorption due to tissue and total blood volume (venous blood and diastolic volume of the arterial blood) are represented with the DC component (Challoner & Ramsay 1974; Lemay et al. 2014). Alterations in the DC level component can be observed due to respiratory rhythm, vasomotor activities, thermoregulation, and motion artifacts (Allen 2007).

A PPG sensor can operate in two different modes: transmission and reflection mode (Figure 4). In transmission mode, the tissue is illuminated on one side and the sensor on the other side captures the light transmitted through it. This mode can be used in ear lobes, index fingers, thumbs, and big toes because the thickness of the tissue allows light transmission (Allen 2007). In the reflection mode, an optical sensor is located next to the source of light, and the reflected and scattered light is measured. Reflection mode PPG sensors can be used on the body, and are typically worn on the hand, wrist, forearm, ankle, forehead or torso (Lemay et al. 2014).

The different optical characteristics of the different layers of human skin involve multiple light interaction processes (scattering, absorption, reflection, transmission and fluorescence) (R. Rox Anderson & John A. Parrish, 1981), all of which affect the PPG signal. The anatomical structure of the human skin consists of three main layers (illustrated in Figure 5), all of which impact on the reflective PPG signal (Shi 2009). Although the epidermis, including the stratum corneaum (100 μm thick) contains no blood vessels, the dead cells are continually being replaced on its surface and the melanocyte cells in this layer produce a dark-brown pigment called melanin, which has an absorbent effect on incoming light (Shi 2009).

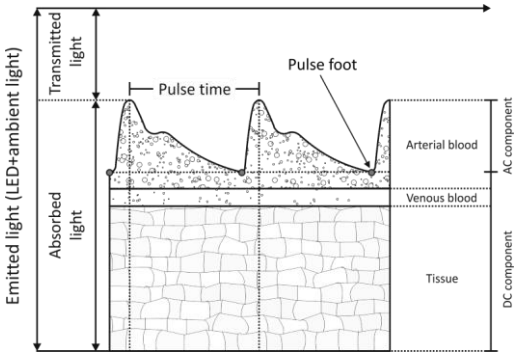


Figure 3: PPG signal components. Reprinted from (Lemay et al. 2014) © 2014 with permission from Elsevier.

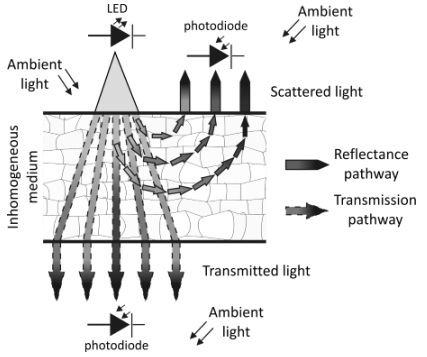


Figure 4: Transmission versus reflectance light PPG modes. Reprinted from (Lemay et al. 2014) © 2014 with permission from Elsevier.

Thus, it can attenuate a PPG signal by decreasing the signal-to-noise ratio from the same-intensity light source. However, the AC component of a PPG signal is not affected by this layer. The dermis (1 – 4 mm thick) contains large networks of arterioles, veinules and capillaries (Shi 2009). This layer produces the AC component of the PPG signal through the scattering effect caused by the interaction between the propagated light and the blood. The third layer, subcutaneous tissue, (1 – 6 mm thick) encloses fat, larger arteries, veins and nerves and is mainly affected by the thermoregulatory functions of the skin and the body. In reflection type sensors, this layer has little effect on the PPG signal because the light is back-scattered in the dermis layer (Shi 2009).

Melanin plays an important role in the optical properties of human skin. The amount of melanin determines the color of the skin, as it is the main absorber of light in the visible spectrum (R. Rox Anderson & John A. Parrish, 1981). The transmittance of skin can thus vary widely between fair- and dark-skinned people (R. Rox Anderson & John A. Parrish, 1981). However, melanin doesn't absorb wavelengths uniformly, as is shown by the graph in Figure 6. It actually absorbs shorter wavelengths better, but at longer IR wavelengths, the absorption of light is almost non-existent (R. Rox Anderson & John A. Parrish, 1981; Lemay et al. 2014). The haemoglobin in red blood cells (40–45% of the cells) also changes the absorption characteristics due to its chemical binding (Shi 2009; Lemay et al. 2014). The solid line in the graph in Figure 6 shows the specific absorption spectra of saturated oxyhaemoglobin (HbO_2) while the dotted line shows the reduced deoxyhaemoglobin (Hb). Arterial blood absorbance, which is measured with two wavelengths of light, red and near infra-red, can be used to estimate the level of oxygen in the blood (Webster 1997).

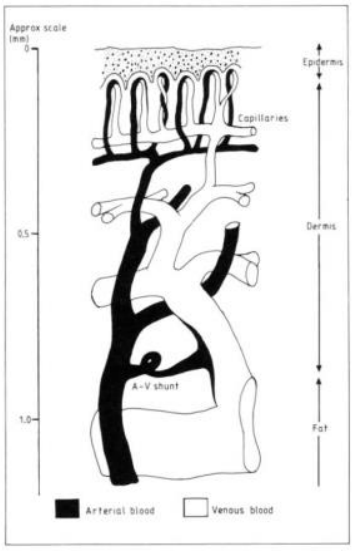


Figure 5: Skin anatomy (Jones 1987 adapted from Conrad 1971). © 1987 IOP Publishing. Reproduced with permission. All rights reserved.

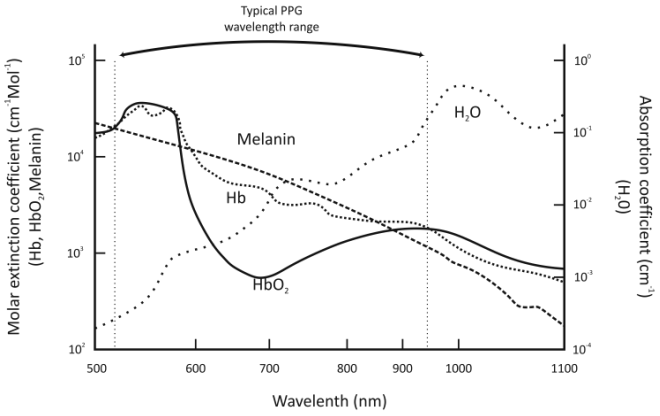


Figure 6: Absorption and molar extinction coefficients of main biological tissue constituents (H2O, Hb, HbO2 and Melanin) at 500 to 1100 nm window wavelengths. Reprinted from (Lemay et al. 2014) © 2014 with permission from Elsevier.

3.3 Main factors affecting PPG signal quality

Being based on the interaction of light and biological tissue, PPG measurements are sensitive to various factors (Lemay et al. 2014). The earliest experimental optical blood flow measurements identified a number of fundamental sources of interference with the measurements: movement and contact of the skin relative to the sensing probe, the size and depth of the vascular area, variation in the intensity and spectrum of the light source used for illumination, and the ratio between reduced and oxygenated hemoglobin on the skin's opacity (Hertzman 1938). The next section describes the main factors impacting on the signal quality of the most recently-designed sensors.

3.3.1 Wavelength and sensor geometry

Both the wavelength of the light source and the sensor geometry, especially the distance between the light emitter and detector, affect the properties and quality of a PPG signal.

In the reflectance mode PPG, the depth to which the light penetrates the tissue (penetration depth) depends on the distance between the LED emitter and the photodetector (PD) receivers (Lemay et al. 2014). The light follows a banana-shaped curve through the tissue from the emitter to the receiver (Reisner et al. 2008). The design and shape of the sensor cause the optical shunting effect, which is the amount of direct light travelling from the emitter to the detector without propagation over pulsing blood in the biological tissue (Webster 1997; Lemay et al. 2014). The optimal distances between a LED source and a PD have been shown to be in the range of 6 to 10 mm for IR light (Mendelson & Ochs 1988) and ~2 mm for green light (Hwang et al. 2016).

Several studies which have analyzed the application of different wavelengths for the PPG light source (in both hot and cold temperatures) have suggested that it is better to use green wavelengths for reflective PPGs. In one study, an examination of the height of reflective PPG signal pulses for finger and forearm signals acquired with four different wavelengths (blue 480 nm, green 560 nm, red 633 nm, and infra-red 825 nm) at 13°C and 42°C peripheral skin temperatures reported the highest AC signal amplitude for the green wavelength in both placement and temperature (Lindberg & Oberg 1991). In another study, a comparison of reflected IR (880 nm) and green (525 nm) PPG measurements performed during rest at 15°C and 25°C peripheral skin temperatures on light-skinned subjects showed over two times higher AC/DC component ratio for the green wavelength, especially at lower temperatures (Maeda et al. 2008). Similar results showing a higher AC/DC component ratio for green wavelength PPG signals than for IR signals have also been observed during rest at normal temperatures, and below 20°C and over 38°C (Maeda et al. 2011). The longer wavelengths in the IR range penetrate deeper into the tissues (R. Rox Anderson & John A. Parrish, 1981) and produce a more complex reflected signal, especially in the presence of motion (Lemay et al. 2014). However, in a cold ambient environment, blood microcirculation decreases due to vasoconstriction, and it is difficult to reach the deeper tissues that have sufficient blood circulation for PPG with shorter wavelengths (Lemay et al. 2014). In addition, dark skin pigmentation (high melanin concentration) had a low reflectance of wavelengths shorter than 650 nm (Cui et al. 1990). Therefore, dark skin or a cold climate may indicate that IR light is preferable. In short, the selection of the right light wavelength depends on the absorbance in the skin, the ambient environment and other factors specific to each use case.

3.3.2 Ambient light

The interference of ambient light is a significant artifact in PPG monitoring. If the PD is exposed to too much ambient light from either a natural or artificial source, its output may

be saturated (Webster 1997). Poor mechanical design, or the incorrect attachment of the device to the body are typical causes. If the ambient light is very bright, e.g. direct sunlight, it may pass through the human skin and tissue and be picked up by the PD through the tissue. Ambient light interference is common and can significantly deteriorate the performance of a PPG monitor unless special care is taken.

The interference of the ambient light can be categorised according to whether it is static or time-varying (Winokur et al. 2015). Static ambient light artifacts are produced by external, high-intensity light sources, such as the sun, which usually increases the DC component of the PPG signal. A high DC component in a PPG signal reduces the dynamic range of the sensor and may saturate an input analog-to-digital convertor. Time-varying ambient light artifacts (e.g. artificial office light) can produce high harmonic frequencies (Winokur et al. 2015). These frequencies can distort and interfere with PPG signals because of PPG's relatively low measurement sampling rates and the aliasing effect. Another common artifact with time-varying ambient light is the "shuttering effect", which is generated by high-frequency changes in the intensity of the ambient light. It occurs when the ambient light level changes rapidly from bright light to shadow and vice versa, e.g. when running in a forest with bright sunlight shining through the trees. The shuttering effect is difficult to filter out, because it is usually modulated in a useful frequency range, i.e. one that is close to that of HR.

The influences of ambient light can be minimized if the sensors are well designed and manufactured. Placing a light filter over the PD (Webster 1997) helps to minimize ambient light interference and improves the final signal SNR by filtering out the longer wavelengths of sunlight. Signal modulation techniques combined with higher sampling frequencies and further digital filtering are typical strategies for reducing the effects of ambient light in PPG measurements (Patterson et al. 2009; Patterson & Yang 2011; Patterson & Yang 2012). Another method for avoiding the interference of ambient light is alternative sampling utilizing a charge redistribution technique (Kim et al. 2015).

3.3.3 Motion artifacts

The largest category of interferents in PPG signal measurements are motion artifacts. Identification and classification of particular motion artifacts in recorded PPG signals varies, not least because motion can come from several different sources at the same time (Lemay et al. 2014). For example, there are inner tissue modifications (e.g. motion of the muscles and tendons, and compression or dilation of the tissues) generated by body movements. There is also the shape of soft tissue (e.g. fat and liquids), which can be

changed by gravity or acceleration and whose changes modify the optical paths of the transmitted or reflected optical signals (Lemay et al. 2014). With PPG measurements, the effect of the binding used to attach the optical probe to the skin can act like a mass spring-system, so that local and global movements of the body change the position of the sensor relative to the skin. Due to the inhomogenous structure of the tissue's surface, these changes can affect the optical paths (Lemay et al. 2014). The amplitude and waveform of a typical optical PPG signal modified by pressure applied to the sensor due to the re-distribution of fluid in the tissue occurs as follows (Lemay et al. 2014). The initial increase of pressure between the sensor and skin augments the pulsating component of the PPG. However, if that pressure rises too far above some threshold value, the pulsating AC component might decrease due to the blood vessels being squashed. Spigulis et al. (Spigulis et al. 2007) demonstrated the effect that a gradual increase in pressure on the probe has on PPG signal shape in a reflective multi-wavelength PPG. At shorter wavelengths (violet 405 nm and green 532 nm), the influence of higher pressure on the probe caused noticeable reductions in both the waveform amplitude and the signal baseline. However, for longer wavelengths (red 645 nm), which penetrate deeper under the skin, only the baseline was reduced.

Motion artifacts can be classified into three categories based on their rhythmicity and frequency of occurrence in typical OHR use-cases (Lemay et al. 2014). First, there are rhythmical motion artifacts, which mostly occur during endurance sport activities (e.g. walking, running, biking, or swimming). Next are rhythmical intermittent motions artifacts, which occur during daily activities (e.g. manual or office work). Lastly are non-rhythmical continuous motions which typically occur during ball games, working out in the gym, or in many other daily activities (e.g. keyboard typing).

The influences of motion artifacts can be minimized if the sensors are well designed and manufactured. Among the mechanical design issues are the use of lightweight measurement devices to decrease the impact of external forces. The friction used when attaching a probe to the hand should compensate for possible displacement from loose sensor bindings. The pressure of probe on hand should not be so high as to cause blood vessel 'clutching' (Lemay et al. 2014). The impact that pressure on the probe and the skin has on the quality of the output PPG signal quality has been studied several times, but the results have been inconclusive (Dresher & Mendelson 2006; Maeda et al. 2013). However, one study showed that the application of higher pressure on the sensing probe improved the sensing signal quality for reflectance pulse oximetry on the forehead (Dassel et al. 1995). The importance of establishing optimal sensor pressure has been highlighted in a study of arterial stiffness using reflective PPG (Grabovskis et al. 2013).

OHR sensors are usually worn on the wrist due to the average user's acceptance and familiarity with such placement (Korhonen et al. 2003). However, a study comparing different sites for a PPG sensor on the body showed that a lateral position on the upper arm is the best compromise between a useful PPG signal amplitude and movement artifacts (Maeda et al. 2011).

In order for OHR data to be reliable, dedicated signal processing methods are required to mitigate the effects that motion has on the signals. Motion artifacts in a recorded PPG signal are suppressed, or at least reduced, during the PPG signal enhancing process. The algorithms used in this process normally get information about the motion from additional motion sensors. Typically, a 3D accelerometer sensor directly measuring propagated motion signals is used for this purpose. Other options include an extra light emitter at a different wavelength with minimal attenuation from the color of the blood. Another, less common, solution is to integrate a pressure sensor into the optical probe (Lemay et al. 2014). The simplest approach is just to discard the segments in which motion artifacts are present, but this makes the measurements somewhat less robust. The more advanced approaches are usually based on the assumption of the stationarity of rhythmical motion artifacts and an additive model, which combines the motion artifacts and the HR components into one optical signal. The spectrums of both the PPG and motion reference signals are estimated and the spectral peaks are identified. Then, the PPG signal is enhanced by filtering out the rhythmical motion frequencies until only the non-motion frequency peaks remain (Lemay et al. 2014). A yet more robust and advanced method, which is not limited only to rhythmical motion artifacts, is adaptive filtering (Haykin 2001). To do this, one needs to find a model that maps the motion signal into the existing components in the PPG signals. Then the motion components are subtracted from the optical signals (Renevey et al. 2001). There are many other inventive methods for enhancing optical PPG signals, such as: an adaptive comb filter with an adaptive IIR Notch Filter structure (B. Lee et al. 2011), adaptive noise cancellation using a normalized least-mean-square algorithm to attenuate motion artifacts and reconstruct multiple PPG waveforms (Fallet & Vesin 2017), a Wiener filter to attenuate the motion artifacts combined with a phase vocoder to refine the HR estimate (Temko 2015), or combining temporally-constrained independent component analysis and adaptive filters to extract clean PPG signals from a motion artifact-corrupted signal (Peng et al. 2014).

3.3.1 User characteristics

User characteristics such as skin color (pigment and melanin), age, or gender do have an impact on PPG waveform morphology, which in turn affects the quality of the measured signal. There are no studies focusing on the effect that other skin properties can have on PPG signal quality, such as hydration, hairiness, sweating or body mass index (BMI)-related parameters.

The influence of five skin types (Fitzpatrick scale I – V. (Fitzpatrick 1988)) at four light wavelengths (blue 470 nm, green 520 nm, red 630 nm, and infrared 880 nm) on reflective PPG signal modulation (AC/DC ratio) was investigated with measurements taken at rest and during exercise (Fallow et al. 2013). The results of this study showed that green light has the best modulation factor at rest regardless of skin type. During exercise, either blue or green had the highest signal-to-noise ratio, depending on the skin type. It was also noted that the darkest skin type (Fitzpatrick class V) produced the poorest quality signal when compared to the other lighter skin types, whether at rest or during exercise. Fallow et al. (Fallow et al. 2013) deduced that this was because melanin affects the light in the epidermal layer of the skin where there are no blood vessels, and so it is a static factor that has the same effect regardless of the conditions. Increased error in HR estimation in subjects with darker skin was also reported during a testing protocol which included a variety of exercises (Spierer et al. 2015). Experimental measurements proved that Caucasian and Asian skin colors have better skin tissue reflectance than dark skins, which have higher pigmentation and melanin concentrations (Cui et al. 1990). The weak light reflection from dark skin pigment can be compensated for by applying a stronger light source (Cui et al. 1990) because the origin of the AC signal component lies beneath the epidermis layer, which doesn't affect the modulation. Another option for reducing the attenuation of the signal from dark skin is to use longer wavelengths close to IR light, as these have better skin penetration (Lemay et al. 2014).

The effects of age are characterized primarily by analyzing the PPG pulse shape. Examination of the PPG pulse shape measured by a reflective probe on three different body parts (fingers, ears and toes) was performed on healthy subjects divided into four age groups (younger than 30 years, 30–39 years, 40–49 years and 50 years or older) (Allen & Murray 2003). The median differences of the normalized PPG pulse shapes between the oldest group and the three younger groups demonstrated evidence of gradual changes with age, particularly a decrease in the amplitude. Significant changes in other parameters affecting the shape of the PPG waveform were observed with increasing age in studies exploring the use of PPG signals to assess arterial stiffness (Brillante et al.

2008; Jayasree et al. 2008; Shi et al. 2009; Wowern et al. 2015). All in all, it appears that age does have an effect on the shape of the PPG waveform. Changes in the PPG waveform, especially the reduction in amplitude, might cause a decrease the AC/DC component ratio and also a decrease in the SNR. Gender differences affecting PPG signal parameters have only been studied in relation to respiratory rate detection in a reflective PPG signal (Nilsson et al. 2006), and no significant difference between the genders was found.

3.3.2 Blood perfusion

An important physiological factor affecting PPG quality is blood perfusion around the sensor area. The PPG signal is particularly sensitive to skin temperature and, more broadly, the whole body temperature. Increases in the AC and DC components of the PPG signal were observed during experimental measurements performed with increasing ambient temperatures from 15°C to 25°C, resulting in the vasodilation effect (Kamal et al. 1989). In a cold environment, typically in a cold room or outdoors, blood perfusion on the skin is reduced to minimize the loss of body heat. This reduces variations in blood volume close to the skin, and in extreme climates the vasoconstriction may cause these variations to disappear altogether. This phenomenon was demonstrated with a comparison of the frequency spectrums of PPG signals from 15°C to 25°C, where the low temperature showed a noticeable decrease of spectral energy in the HR frequency (Kamal et al. 1989). A comparison of PPG signal measurements recorded in different seasons and ambient room temperatures also reported variations in the PPG signal quality, and there is a significant decrease in PPG amplitude during the cold winter season (Kumazawa et al. 1964). Cold conditions reduce the SNR and may cause more artifacts, or even a complete loss of the optical signal. Obviously, wearing warm clothes to keep the skin warm around the sensor will help to alleviate the problem, as will strenuous physical activity which produces energy and heat which raises the body temperature and improves blood perfusion.

Mental stress has also been shown to have an impact on a PPG signal (Kumazawa et al. 1964). One of the first PPG studies showed that acute mental stress causes general vasoconstriction in the peripheral vessels and thus reduces the amplitude of a PPG signal. When the subject grew accustomed to the test, the stress effect was reduced and the signal returned to normal values. This shows that the effect is probably more related to emotional excitement than to the cognitive process of thinking itself (Kumazawa et al. 1964). Thus, stress situations might cause a deterioration of the SNR in a PPG signal.

3.4 Estimation of VO_{2max} and energy expenditure from heart rate

The main determinants of athletics endurance performance are the following: maximal oxygen uptake (VO_{2max}), lactate threshold, running economy, fractional utilization of VO_{2max}, and speed during maximal anaerobic test (Paavolainen et al. 1999; Bassett & Howley 2000). The VO_{2max} parameter represents cardiorespiratory endurance capacity or aerobic power which is capacity of the body to distribute and utilize oxygen during maximal exertion involving dynamic contractions of large muscles (Nieman 2011). The VO_{2max} measurement is important to describe cardiorespiratory fitness of individuals to examine possible risk of premature death from all causes, especially heart diseases (Nieman 2011). VO_{2max} can be estimated directly in combination of measuring ventilation and analyzing the amount of air exhaled in the form of carbon dioxide in the exhaled breath compared to the amount of oxygen in the inhaled breath (Franklin & Balady 2000; Nieman 2011). The standard laboratory method of estimating VO_{2max} is to do graded aerobic maximal exercise, typically running until the point of total exhaustion. At the point of total exhaustion, the subject has reached both VO_{2max} and the maximal attainable HR (Nieman 2011). The laboratory method is, of course, impractical for monitoring any uncontrolled activities. However, VO_{2max} can also be estimated during submaximal exercise. This is achieved by utilizing the linear relationship between HR, oxygen uptake and workload in combination with the maximum age-based HR calculation (Nieman 2011). Reis et al. (Reis et al. 2011) used this method to monitor well-trained long-distance runners on a running track. They reported a very high linear regression between oxygen uptake and HR; HR and running velocity; and, oxygen uptake and running velocity. They concluded that HR can be used to predict the energy demand for a specified running speed.

Metabolic processes, including exercise, are generating heat which is directly proportional to energy expended (Franklin & Balady 2000). The estimation of energy expenditure (EE) can be used to determine the effect of physical activity (Franklin & Balady 2000). Total energy expenditure (TEE) has three components: basal metabolic rate (BMR), the thermic effect of food and the EE of the activity (Levine 2005). There are three different approaches to measuring EE. The most common method used in sport or research is indirect calorimetry (IC) performed with gas exchange analyses (measurement of oxygen consumption and carbon dioxide production) (Levine 2005; Nieman 2011). Direct calorimetry is based on measuring the subject's heat loss rate, which can be done in calorimetric chambers or with heat suits (Levine 2005). The non-calorimetric methods usually

use physiological markers, such as specific hydrogen isotope dilution in the doubly labeled method, or the measurement of other physiological variables (Levine 2005). Keytel et al. (Keytel et al. 2005) proposed and validated mixed-model equations with user characteristic parameters (gender, height), both with and without the fitness level parameter (VO_{2max}) for prediction of EE from HR during physical activity. Their results confirmed a strong agreement between predicted EE and reference EE measurement utilising IC and both models for EE prediction (with and without VO_{2max}). Charlot et al. (Charlot et al. 2014) improved the models proposed by Keytel et al. by adding actual running speed, resting HR, speed at VO_{2max} or substituting HR with speed. It was concluded that those models that included running speeds provided the most accurate EE and the closest agreement with the IC method. Altini (Altini 2015) presented models and methods to provide accurate EE and VO_{2max} estimation for individuals without requiring individual calibration. His protocol was in free-living conditions with wearable sensors measuring HR and inertial accelerometry.

4 Evaluation of wearable optical heart rate monitors

Several studies aimed at evaluating the accuracy of ECG-based consumer-wearable HR monitors have been published since the mass production of such devices began in the 1980s. The early studies focused on evaluating chest-strap HR monitors in sports activities, while later studies also included their use in other applications. Several evaluation studies of OHR monitors have been published since 2016 but the main topic of interest is still the accuracy of OHR during sport activities. Until very recently, very few studies have evaluated other applications of OHR. Several studies have examined the accuracy of VO_{2max} and EE estimations on the basis of measured HR, including measurements from OHR. The first section of this chapter describes the methods and statistical error metrics and then applies them to the estimations of HR and IBI accuracy. The methodology and results of the most relevant evaluation studies are then summarised.

4.1 Methods and metrics used for evaluating the accuracy of wearable heart rate monitors

Various statistics and visual representations have been used to demonstrate the accuracy of wearable HR monitors. Typically, slightly different statistics are needed to assess the accuracy of HR (in bpm) or HRV (RRI or IBI in ms). The most commonly used statistical methods are summarized in Table 1. The metrics listed in Table 2 can all be used to estimate the accuracy of the HR measurements. The relative or absolute number of missing, correct, and extra detected beats can also be calculated using the automatic method (Parak et al. 2015; Pietilä et al. 2018). This method is based on checking the number of corresponding reference beats to one detected beat from the tested device within a defined range [$t - p \cdot l$, $t + p\ l$], where t is the time when the beat was detected, p is a parameter for limiting the search range, and l is the length of IBI (example in Figure

7). In this case, the parameter *p* was empirically set up at 0.5, which corresponds to half of the IBI length. The beats at positions *t* = 1000 ms, 2000 ms and 4500 ms are properly detected because they have only one corresponding reference beat. For the beat at position t = 3050 ms, there are two corresponding reference beats, so it is assumed that in this case a beat was missed. For the beat at position t = 5500 ms, there is no corresponding reference beat, so this is considered to be an extra beat.

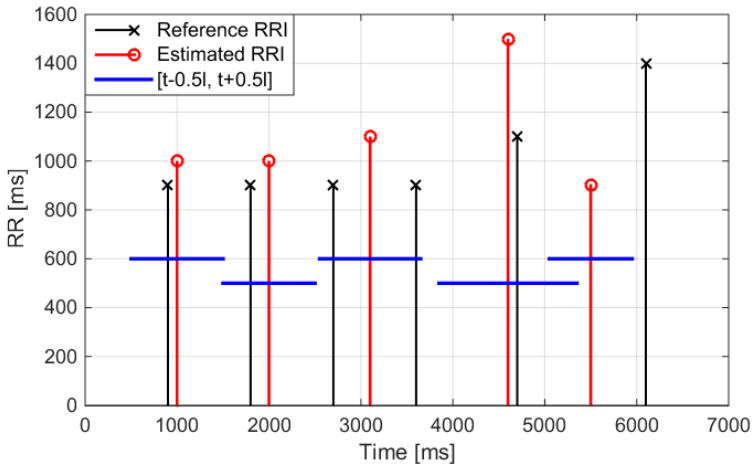


Figure 7: Illustrative example of detecting extra and missing beats (Parak et al. 2015). © 2015 IEEE. Reprinted with permision.

The Bland–Altman plot (BA – Plot) is a visualization method used to compare two different measurement approaches (Altman & Bland 1983; Bland & Altman 1986), such as the results from the tested device and from the reference device (Examples in Figure 8). In this graphic method, the differences between the pair values are plotted against the averages of the pair values. The graph therefore provides a good illustration of the differences in the application of the two different measurement methodologies. The plot also contains horizontal lines signifying 95% limits of agreement (LoA) (mean difference ± 1.96 SD of differences) and the mean difference, which describes the bias (systematic measurement error) between the measurement methods. The BA-Plot can be extended by displaying the distributions of the mean and the difference of the values (Figure 8b)

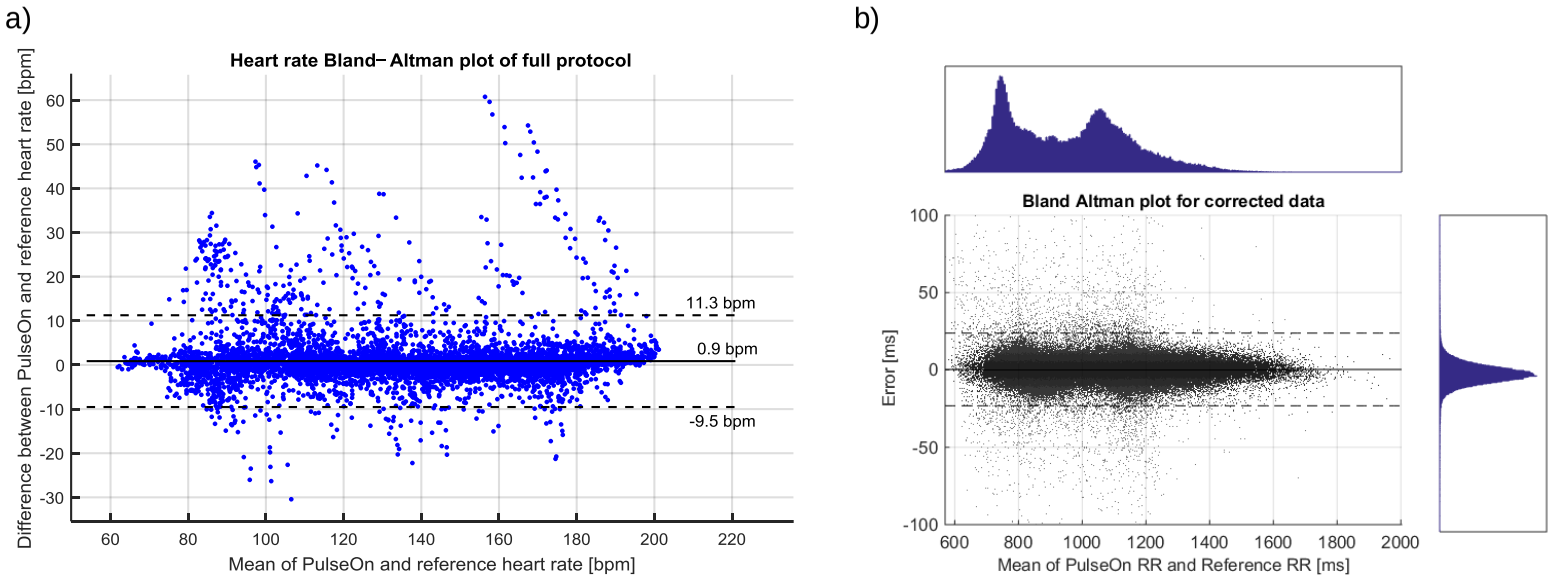


Figure 8: Bland-Altman plots examples a) HR during test running protocol (Parak et al. 2017). © 2017 JMIR mHealth and uHealth. Published and reproduced under terms of Creative Commons License 4.0. b) IBI during sleep recording (Parak et al. 2015). © 2015 IEEE. Reprinted with permision.

Table 1: Summary of the general statistical methods and error metrics for estimation of accuracy

Name	Abbreviation	Unit	Description	Equation
Mean error	ME	[unit of variable]	Average error of all dataset errors	$ME = \frac{\sum_{i=1}^{n} x_i - y_i}{n}$
Mean absolute error	MAE	[unit of variable]	Average of all absolute dataset errors	$MAE = \frac{\sum_{i=1}^{n} \lvert x_i - y_i \rvert}{n}$
Mean percentage error	MPE	[%]	Average of all percentage dataset errors	$MPE = \frac{100\%}{n} \sum_{i=1}^{n} \frac{x_i - y_i}{y_i}$
Mean absolute percentage error	MAPE	[%]	Average of all absolute percentage dataset error	$MAPE = \frac{100\%}{n} \sum_{i=1}^{n} \left\lvert \frac{x_i - y_i}{y_i} \right\rvert$
Pearson's correlation coefficient	*r*	[-1, 1]	Measure of strength and direction of a linear relationship between two variables	$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$
Spearman's rank correlation coefficient	(Rho) ρ	[-1, 1]	Non-parametric measure of statistical dependence between the ranking of two variables	$\rho = \frac{cov(rg_x rg_y)}{\sigma_{rg_x}\sigma_{rg_y}}$
Standard error of estimate	SEE	[-]	Measure of the accuracy of predictions made with a regression line	$SEE = \sqrt{\frac{\sum_{i=1}^{n}(x_i - y_i)^2}{n}}$
Mean value and standard deviation	Mean ± SD	[unit of variable]	Average value and standard deviation for whole measurement for specific dataset	

[1] The variables in the equations are: n is the number of samples, x_i are the tested device samples, y_i are the reference device (true) samples, $\bar{x}$ is the sample mean , and analogously for $\bar{y}$, $cov(rg_x rg_y)$ is the covariance of rank variables, $\sigma_{rg_x}\sigma_{rg_y}$, which are the standard deviations of the rank variables.

Table 2: Special HR measurement accuracy metrics

Name	Unit	Description	Equation
HR Score < 5 bpm [%] HR Score < 10 bpm [%]	[%]	% of time when absolute difference between reference and tested device is smaller than 5 bpm or 10 bpm	$HR_{score<p\ bpm} = \frac{100\%}{n}\sum_{i=1}^{n} f_{absscore}(x_i, y_i)$ $f_{absscore}(x,y) = \begin{cases} 1 & \lvert x-y\rvert < p \\ 0 & \lvert x-y\rvert \geq p \end{cases}$
HR Score < 5% bpm [%] HR Score < 10% bpm [%]	[%]	% of time when absolute percentage difference between reference and tested device is smaller than 5% bpm or 10% bpm	$HR_{score<p\%\ bpm} = \frac{100\%}{n}\sum_{i=1}^{n} f_{perscore}(x_i, y_i)$ $f_{perscore}(x,y) = \begin{cases} 1 & 100\%\left\lvert\frac{x-y}{y}\right\rvert < p\% \\ 0 & 100\%\left\lvert\frac{x-y}{y}\right\rvert \geq p\% \end{cases}$
HR Reliability [%]	[%]	% of time when absolute difference between reference and tested device is smaller than 10 bpm, same as HR score < 10 bpm	$HR_{reliability} = HR_{score<10\ bpm}$
HR Accuracy [%]	[%]	Defined as complement of the relative error (i.e. 100% - mean absolute percentage error)	$HR_{accuracy} = 100\% - MAPE$

[2] The variables in the equations are: n is the number of samples, x_i are the tested device samples, y_i are the reference device (true) samples, and p is the acceptance threshold parameter.

4.2 Performance evaluation of ECG-based chest-strap HR monitors

There have been numerous studies to evaluate the performance of consumer chest-strap HR monitors. Some key evaluation studies are summarized in Table 3 in order to provide a benchmark for the consumer HR monitor performance and evaluation methodology. It is noticeable that the studies lack a common evaluation methodology, so this complicates any comparison of the reported results. The selection of subjects, the testing protocol, the signal pre-preprocessing techniques and the statistical methods are not standardised. In particular, inappropriate methods of reading data and its synchronization between devices highlight relevant discrepancies in the studies. Manual reading of the average HR values from a device display and printing the ECG curve for short time period (10 seconds) during the last minute of a performed activity was used in two studies (Terbizan et al. 2002) and (C. M. Lee & Gorelick 2011). The synchronization of the signal between devices was handled with manual alignment of the paired RRI series (Vanderlei et al. 2008; Porto & Junqueira 2009; Weippert et al. 2010) or by utilizing temporal event markers with corresponding time coordinates (Kingsley et al. 2005). None of the studies aimed at evaluating the accuracy of RRI detection took into account any possible inaccuracies in the internal device clock. In addition, two chest straps were placed on the body simultaneously (Weippert et al. 2010), which might cause inaccuracy due to poor contact between the strap and the skin. The impact of placing two chest strap simultaneously has not been systematically tested. The most commonly applied performance evaluation methods for the tested and reference devices in these studies were a graphical representation bias and limits of agreements (LoA) in BA plots, and estimation of the correlation coefficients. Five studies examined RRI estimation accuracy and two studies evaluated HR measurement only. Most studies have focused on sports activities (typically endurance training sports such as running, walking or biking), so evaluations of everyday use have been scarce. The number of participants varies between 8 and 33 with often limited population characteristics (age, gender). In general, the studies show good agreement and low HR estimation error between chest-strap HR monitors and the ECG reference values during sport, with strong correlations during rest $r = 0.83$ — 0.97, walking $r = 0.82$ — 0.97, jogging $r = 0.87$ – 0.9. Accuracy during running, however, appears to be significantly reduced by very weak to strong correlation $r = -0.26$ – 0.81. Chest-strap HR monitors provide good RRI estimation accuracy during rest with corresponding bias = -1.85 – 0.24 ms and LoA varies in the ranges [-6.37 -1.47] ms and [1.96 2.67] ms. The overall results of the sport protocol compared to rest monitoring show a lower bias = 0.41 – 0.44 ms but an increased LoA variation in the ranges [-15.1 -12.4] ms

and [11.5 14.3] ms. Long-term 24-hours RRI estimation accuracy was examined during a non-controlled protocol including common daily activities (Kristiansen et al. 2011). The results were calculated using the Deming regression method (Linnet 1993) (intercept = 1.91 and slope = 0.998) and therefore they are not directly comparable with the other studies. There were no systematic differences in HRV between the tested device and the reference in this study.

Table 3: Summary of ECG based HR monitors evaluation studies

Study	**Tested devices (M = number of tested chest strap or disposable electrodes HR monitors, devices' names)**	**Reference device type**	**Subjects (N = number of subjects, F = females, age mean ± SD)**	**Protocol description**	**Statistical metrics (desc. in Table 1)**	**Key results (*r* [-1, 1], bias [ms], LoA [ms])**
(Terbizan et al. 2002)	M = 5, Polar Vantage XL, Polar Accurex II, Cardiochamp, Accumen Basix, Cardiosport ZW-8	Holter ECG	N = 14, F = 0, age 19.6 ± 2.3	controlled, sport (rest, walking, jogging, running), total duration 40 minutes	*r*, SEE	HR: during whole protocol Polar Vantage XL *r* = 0.81 – 0.95; Polar Accurex II *r* = -0.26 – 0.95; Cardiochamp *r* = -0.64 – 0.92; Accumen Basix *r* = -0.23 – 0.78; Cardiosport ZW-8 *r* = 0.38 - 0.83
(Kingsley et al. 2005)	M = 1, Polar S810	Ambulatory ECG	N = 8, F = 2, age 26.6 ± 3.6	controlled, sport (ergo-cycle pre-exercise, ergo-cycle incremental load), 3 minutes + 2 minutes for each load until volitional fatigue	*r*, BA Plot	RRi: during whole protocol *r* = 0.927 – 0.998, bias <= 0.10 ms
(Vanderlei et al. 2008)	M = 1, Polar S810	Ambulatory ECG	N =15, F = 0, age = 20 ± 1.4	controlled, sport(rest, ergo-cycle), total duration 40 minutes	*r*	RRi: during rest *r* = 1.0 , ergo-cycle *r* = 0.942 – 0.993
(Porto & Junqueira 2009)	M = 1, Polar S810	Ambulatory ECG	N = 33, F = 18, age 26.1 ± 7.8	controlled, rest(supine and standing), total duration 10 minutes	BA Plot	RRi: during supine bias = -1.85 ms, LoA -6.37 to 2.67 ms, during standing bias = 0.24 ms, LoA -1.47 to 1.96 ms
(Weippert et al. 2010)	M = 2, Polar S810, Suunto t6	Ambulatory ECG	N = 19, F = 0, median age 24	controlled, sport(rest, walking, limbs exercise), total duration 40 minutes	BA Plot	RRi: during whole protocol Polar S810 bias = 0.41 ms , LoA 15.1 to 14.3 ms; Suunto t6 bias = 0.44 ms, LoA 12.4 to 11.5 ms
(C. M. Lee & Gorelick 2011)	M = 1, Polar Vantage XL	Ambulatory ECG	N = 25, F = 13, age 27.3 ± 5.2	controlled, sport(rest, walking, jogging), total duration 18 minutes	*r*, SEE	HR: during rest *r* = 0.97, walking *r* = 0.95 – 0.97, jogging *r* = 0.98
(Kristiansen et al. 2011)	M = 1, Actiheart	Holter ECG	N = 8, F = 8, age 35.50 ± 17.36	non-controlled, daily activity (physical work, leisure time, sleep), total duration 24 hours	Deming regression analysis, BA Plot	RRi mean: intercept = 1.91, slope = 0.998. No systematic differences between Actiheart and Holter HRV were found.

4.3 Performance evaluation of PPG-based consumer OHR monitors

Although the performance of consumer OHR monitors has lately been the subject of rigorously objective study (Table 4 and Table 5), the design of the evaluation protocols, the selection and number of subjects, and the evaluation metrics vary widely between the studies. As the results for chest strap HR monitors show, it can be difficult to compare the outcomes of studies. For example, the number of participants in the studies varies from 10 to 68, and although most of the studies typically involve sport activities (14 studies), there has also been research into the devices' performance in everyday life (4 studies), sleep (2 studies), and there have been two in a clinical environment. Two studies have evaluated the beat detection and IBI estimation accuracy of wrist-worn OHR devices. Among of all of the studies, the most common performance evaluation metrics were Bland-Altman plots (including bias and LoA), estimation of the correlation coefficient, mean absolute error (MAE) or mean absolute percentage error (MAPE). In several studies, the devices' performance was also evaluated with a t-test comparison of the mean difference of the average HR values (Valenti & Westerterp 2013; Olenick et al. 2015; Spierer et al. 2015; de Zambotti et al. 2016; Boudreaux et al. 2018). This method shows if the mean difference between the tested and reference devices is statistically significant. However, it is not ideal for our purposes for a number of reasons. First, it does not measure the amount of differences between methods. In addition, an increasing of number of samples makes the test more sensitive to the difference between the tested and reference datasets. Moreover, if the comparison uses the average HR over a long time-period (or over several activities), this can mask any inaccuracy in the tested device during any one activity. Three of the evaluation protocols had more than one device placed on the wrist (Stahl et al. 2016; Shcherbina et al. 2017; Boudreaux et al. 2018). However, using more than one device at once on one hand can distort signal quality, and this can cause significant inaccuracy in the HR estimation. Synchronization between devices was performed by cross-correlating acquired HR data (Hwang et al. 2016; Pietilä et al. 2018) or manual timestamp-based alignment (Kroll et al. 2016; Jo et al. 2016; J. Lee et al. 2016). The Viterbi (Viterbi 1967) algorithm was applied to align IBI series in order to compensate for the differences between the offset and clock drifts of different measurement systems (Renevey et al. 2013). The measurements were usually split into smaller segment to obtain more accurate statistics for particular activities in the protocols. The duration of the HR segments used as inputs for the statistics gathered in Table 4 and Table 5 varied from 5 seconds to 5 minutes, in addition to which, the methodologies

for the synchronization and segmentation of the HR signals were not available for several of the studies (Olenick et al. 2015; Stahl et al. 2016; Wallen et al. 2016; de Zambotti et al. 2016; Dooley et al. 2017; Shcherbina et al. 2017; Khushhal et al. 2017; Boudreaux et al. 2018) which adds to the difficulty of interpreting and comparing all the reported results.

During sport activities, the results of the studies on OHR performance varied widely; MAPE 1.14 – 25.38%, strong correlation $r = 0.81 – 0.97$, bias varies from -9.3 to 12.0 bpm with a relatively large LoA variation in the ranges [-41.84 -7.0] to [7.3 42.0] bpm. This wide range in performance might be caused by the different activities involved in the testing protocols (i.e. rest, running, cycling, limb exercises), and of course, the devices themselves, because the performance and accuracy of different OHR monitors varies widely for a number of reasons, such as the tightness of the bindings, the design of the sensor, or the algorithms used. However, daily activities usually include more non-rhythmic movements as a source of possible signal artifacts. The error range in MAPE was 1.84 – 9.71% in HR estimation. There was very good HR estimation accuracy represented by very small mean bias; (95% LoA) 0.88 (0.04, 1.72) bpm was measured during sleep for healthy adolescents (de Zambotti et al. 2016). Similar results, i.e. a strong correlation $r = 0.99$ and small bias (95% LoA) -0.05 (-2.454 to 2.43) bpm were reported during open and laparoscopy pediatric surgery (Pelizzo et al. 2018). Moderate correlation $r = 0.74$ and higher mean bias 4.7 ms with wider limits of agreement (95% LoA) (-31 to 21) bpm were reported during 24 hour-measurements in an intensive care unit testing patients with sinus and non-sinus rhythm (Kroll et al. 2016). In another trial, during 8-hour daily activities measurements were taken to test the difference in the HR measurement error between the non-dominant and the dominant hand, and the devices scored MAPE 9.17% vs. 9.71% respectively (J. Lee et al. 2016). Clinically validated OHR has shown very good reliability (HR Score < 10 bpm) 93.8% and low estimation error MAE 3.1 bpm, MAPE 3.1% (Hendrikx et al. 2017). Although some evaluation studies provided skin color classification in the subjects' demographic description (Spierer et al. 2015; Hendrikx et al. 2017; Wang et al. 2017; Shcherbina et al. 2017; Khushhal et al. 2017), it was not mentioned in the reports of the key results. However, an increase in relative mean error associated with darker skin subjects was observed for Mio Alpha OHR during sport testing (Figure 9) (Spierer et al. 2015).

The number of studies evaluating IBI estimation with a wrist OHR is still limited. The beat-to-beat estimation accuracy of the PPG wristwatch system and the gold standard ECG-based RRI during full night polysomnography showed an overall good agreement between both approaches at 0.05 ms ± 18 ms (mean ± SD) and (95% LoA) (-35.7 to 35.81) respectively (Renevey et al. 2013). The relative amounts of detected beats during

a daily protocol for two consumer wrist OHR devices, the PulseOn and Empatica E4, against an ECG based RRI recorder were 76.2 – 90.3% and 9.1 – 67.9%, respectively (Pietilä et al. 2018). For both devices, the smallest number of correctly detected beats occurred during the household chores part of the protocol, which mostly consisted of non-rhythmic movements. The devices' opto-mechanical design and the applied algorithm might explain the significant difference in beat detection accuracy between the PulseOn and Empatica E4 devices. Schäfer and Vagedes (Schäfer & Vagedes 2013) provided a comprehensive review comparing IBI and RRI estimation between different experimental finger PPG-based systems and standard ECG. The studies described in the review were typically performed at rest (sitting, standing or supine position) and reported high agreement between mean IBI pairs (average over few minutes) and a consistently strong correlation, r = 0.97 – 0.99 and mean bias, 0.01 – 0.1 ms. Although it is difficult to draw quantitative conclusions because of the different testing methodologies, it can be concluded that PPG-based IBIs are accurate enough to estimate the HRV for a healthy person at rest. There were no conclusive findings about either the position of the sensor or the ideal detection algorithm.

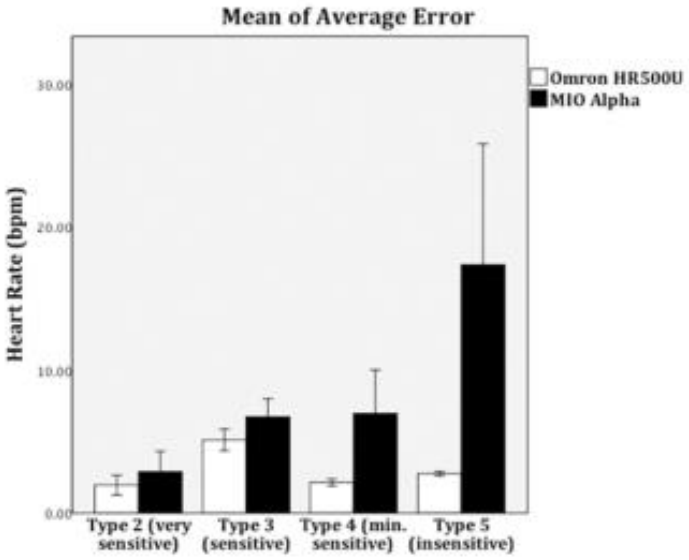


Figure 9: Mean error related to skin color type according Fitzpatrick scale (Spierer et al. 2015). © 2015 Reproduced with permission from Taylor & Francis.

Table 4: Summary of PPG based HR monitors evaluation studies

Study	**Tested devices (M = number of tested devices, devices' names)**	**Reference device type, name**	**Subjects (N = number of subjects, F = females, age mean ± SD, skin color if available)**	**Protocol description**	**Statistical metrics (according Table 1)**	**Key results (*r* [-1, 1], bias [bpm], LoA [bpm], MAE [bpm], MAPE [%])**
(Valenti & Westerterp 2013)	M = 1, prototype of Philips Optical Heart Rate Module	Holter ECG	N = 24, F = 10, age 28 ± 9, subjective skin color classes: 1 dark (African), 5 medium (Asian), 18 white (Caucasian)	controlled, sport (rest supine and standing, walking, running, rest sitting), total duration 42 minutes	ME ± SD of ME, t-test	HR: during whole protocol ME ± SD of ME 0.1 ± 0.3 bpm
(Olenick et al. 2015)	M = 1, Mio Alpha	Holter ECG	N = 10, F = 4, age 35.8 ± 7.8	controlled, sport (rest, running according Bruce Protocol graded exercise test until volitional fatigue), total duration N/A	*r*, mean ± SD HR, t-test	HR: during whole protocol *r* = 0.98, mean ± SD: ECG reference = 122 ± 38 bpm, Mio Alpha = 122 ± 37 bpm
(Kroll et al. 2016)	M = 1, Fitbit Charge HR	Clinical ECG, BedMasterEX ICU bed side monitor	N = 50 (42 with sinus rhythm, 8 with no-sinus rhythm), F = 24, mean age 64	controlled, clinical (Intensive care units patients), total duration 24 hours	*r*, BA Plot	HR: during whole protocol *r* = 0.74, bias = 4.7 bpm LoA -31 to 21 bpm
(Spierer et al. 2015)	M = 2, Omron HR500U, Mio Alpha	Chest strap ECG, Polar RS800CX	N = 47, age 28.4 ± 10.1, Fitzpatrick skin type 3.0 (SD = 0.7)	controlled, sport (rest, walking, jogging, elliptical, stairs climbing, ergo-cycling, weights lifting), total duration 48 minutes	ME ± SD of ME, t-test	HR: during whole protocol ME ± SD of ME was for Mio Alpha from 2.39 ± 6.28 bpm to 23.30 ± 31.94 bpm; Omron HR500U from 2.22 ± 3.67 bpm to 4.67 ± 8.95 bpm
(Stahl et al. 2016)	M = 5, Scosche Rhythm, Mio Alpha, Fitbit Charge HR, Basis Peak, Microsoft Band, TomTom Runner Cardio	Chest strap ECG, Polar RS400	N = 50, F = 18, age males 27.47 ± 6.1, age females 24.17 ± 4.93	controlled, sport (rest, walking, running), total duration 36 minutes	*r*, MAPE, BA Plot	HR: during whole protocol Scosche Rhytm *r* = 0.93, MAPE = 3.98%; Mio Alpha *r* = 0.93, MAPE = 4.56%; TomTom Runner Cardio *r* = 0.96, MAPE = 3.28%; Microsoft Band *r* = 0.96, MAPE = 4.82%; Basis Peak *r* = 0.95; Fitbit Charge HR *r* = 0.93, MAPE = 6.22%

Table 4: continued

Study	Tested devices (M = number of tested devices, devices' names)	Reference device type, name	Subjects (N = number of subjects, F = females, age mean ± SD, skin color if available)	Protocol description	Statistical metrics (according Table 1)	Key results (*r* [-1, 1], bias [bpm], LoA [bpm], MAE [bpm], MAPE [%])
(Wallen et al. 2016)	M = 4, Apple Watch, Fitbit Charge HR, Samsung Gear S, Mio Alpha	Holter ECG, GE Healthcare CASE exercise testing system	N = 22, F = 11, age 24 ± 5.6	controlled, sport(supine, sitting, standing, running, sitting, ergo-cycle sitting), total duration 58 minutes	*r*, Rho, BA Plot	HR: during whole protocol Apple Watch *r* = 0.95, bias = -1.3 bpm, LoA -9.9 to 7.3 bpm; Fitbit Charge HR *r* = 0.81, bias = -9.3 bpm, LoA -26.0 to 7.4 bpm; Samsung Gear S ρ = 0.67, bias = -7.1 bpm, LoA -27.3 to 13.1 bpm; Mio Alpha *r* = 0.87, bias = -4.3 bpm, LoA N/A
(de Zambotti et al. 2016)	M = 1, Fitbit Charge HR	Clinical polysomnography ECG	N = 32 (adolescents), F = 12, age 17.3 ± 2.5	controlled, sleep, (first night for polysomnography adaption, then next night with polysomnography and tested device), total duration N/A	BA Plot, mean ± SD, t-test	HR: during whole protocol bias = 0.88 bpm, LoA 1.72 bpm to 0.04 bpm; mean ± SD: ECG reference 60.2 ± 7.6 bpm, FitBit charge HR 59.3 ± 7.5 bpm
(Jo et al. 2016)	M = 2, Basis Peak, Fitbit Charge HR	Ambulatory ECG, Cosmed C12x	N = 24, F = 12, age 24.8 ± 2.1	controlled, sport (rest, ergo-cycle, walking, jogging, raise with dumbbells, isometric plank), total duration 77 minutes	*r*, BA Plot	HR: during whole protocol Basis Peak *r* = 0.92, bias = -2.53 bpm LoA -24.37 to 19.31 bpm; Fitbit Charge HR *r* = 0.83, bias = -8.79 bpm, LoA -41.84 to 24.24 bpm
(J. Lee et al. 2016)	M = 2, Fitbit Charge HR on dominant hand, Fitbit Charge HR on non-dominant hand	Chest strap ECG, Polar RS400	N = 10, F = 2, age 26.5 ± 5.4	non-controlled, daily activity, total duration 8 hours	*r*, MAPE	HR: during whole protocol Fitbit Charge HR on non-dominant *r* = 0.80, MAPE = 9.17%; on dominant hand *r* = 79, MAPE = 9.71%
(Hwang et al. 2016)	M = 1, Basis Peak	Chest strap ECG, Polar H7	N = 11 (manual workers), age 26 ─ 60	non-controlled, daily activity (rest, manual working), duration from 70 to 360 minutes for subject, in total 1420 minutes of data	*r*, MAPE, BA Plot	HR: during rest *r* = 0.92, MAPE = 4.07%, bias = 1.56 bpm , LoA -9.10 to 12.22 bpm; during manual working *r* = 0.77, MAPE = 4.96%, bias = 2.64 bpm, LoA -10.66 to 15.95 bpm; whole protocol *r* = 0.85, MAPE = 4.79%, bias = 2.30 bpm, LoA -10.71 to 15.31 bpm

Table 4: continued

Study	Tested devices (M = number of tested devices, devices' names)	Reference device type, name	Subjects (N = number of subjects, F = females, age mean ± SD, skin color if available)	Protocol description	Statistical metrics (according Table 1)	Key results (r [-1, 1], bias [bpm], LoA [bpm], MAE [bpm], MAPE [%])
(Hendrikx et al. 2017)	M = 1, Philips Health Watch	Holter ECG, Actiware Cardio and Chest strap ECG, Polar T34	N = 29, F = 15, age 41.2 ± 14.4, Fitzpatrick skin types (1 – 6) distribution: 0, 7, 18, 4, 0, 0	controlled, sport (rest, walking, ergo-cycle, cross-trainer, household chores, office work, lying, standing, walking, cycling, running), total duration 75 minutes	ME, MAE, MPE, MAPE, HR score < 10 bpm, HR score < 10% bpm	HR: during whole protocol ME = -1.7 bpm , MAE = 3.1 bpm, MPE = -1.3%, MAPE = 3.1%, HR score < 10 bpm = 93.8%, HR score < 10% bpm = 93.1%
(Wang et al. 2017)	M = 4, Apple Watch, Mio Fuse, Fitbit Charge HR, Basic Peak	Ambulatory ECG	N = 50, F = 28, age 37 ± 11.3, 7 black skin African-American	controlled, sport (rest, walking, jogging, running, recovery) total duration 20 minutes	concordance correlation coefficient (r_c)	HR: during whole protocol Apple Watch r_c = 0.91, Mio Fuse r_c = 0.91, Fitbit Charge HR r_c = 0.84, Basis Peak r_c = 0.83
(Dooley et al. 2017)	M = 4, Apple Watch, Fitbit Charge HR, Garmin Forerunner 225	Chest strap ECG, Polar T31	N = 62, F = 36, age 22.55 ± 4.34	controlled, sport(rest, walking, jogging, recovery), total duration 40 minutes	*r*, MAPE, BA Plot, mean ± SD HR	HR: during whole protocol Apple Watch MAPE = 1.14 – 6.70%; Fitbit Charge HR MAPE = 2.38 – 16.99%; Garmin Forerunner 225 MAPE = 7.87 – 24.38%
(Claes et al. 2017)	M = 1, Garmin Forerunner 225	ECG Recorder, Zensor	N = 12, F = 6, mean age 28	controlled, sport (standing, sitting, walking, rest periods), total duration 40 minutes	*r*, BA Plot, RMSE	HR: during whole protocol *r* = 0.65 – 0.868, bias = 1.57 bpm, LoA 32.53 to 29.40 bpm, RMSE = 3.01 bpm or 2.89%.
(Shcherbina et al. 2017)	M = 6, Apple Watch, Basis Peak, Fitbit Surge, Microsoft Band, Mio Alpha 2, PulseOn, Samsung Gear S2.	Ambulatory ECG	N = 60 , F = 31, age 38 ± 11, Fitzpatrick skin type: males 3.7, (SD = 1.39), females 3.7 (SD = 1.25)	controlled, sport(rest, walking, running, ergo-cycle), total duration 40 minutes)	MPE	HR: visual representation of MPE for specific activities, numerical results N/A

Table 4: continued

Study	Tested devices (M = number of tested devices, devices' names)	Reference device type, name	Subjects (N = number of subjects, F = females, age mean ± SD, skin color if available)	Protocol description	Statistical metrics (according Table 1)	Key results (r [-1, 1], bias [bpm], LoA [bpm], MAE [bpm], MAPE [%])
(Benedetto et al. 2018)	M = 1, Fitbit Charge 2	Ambulatory ECG	N = 15, F = 8, age males 31 ± 4, age females 32 ± 4	controlled, sport (ergo-cycle), total duration 10 minutes	BA Plot	HR: during whole protocol Fitbit Charge 2 bias = -5.9 bpm LoA -28.5 to 16.8 bpm
(Pelizzo et al. 2018)	M = 1, Fitbit Charge HR	Clinical ECG, Infinity Delta, Dräger bed side monitor	N = 30, F = 14, age 8.21 ± 3.09	controlled, clinical (laparoscopy and open surgery patients), total duration 1 hour	*r*, BA Plot	HR: during Fitbit Charge HR whole protocol r = 0.99, bias = -0.05 bpm LoA -2.454 to 2.43 bpm
(Khushhal et al. 2017)	M = 2, Apple Watch on right and left hand	Polar S810i	N = 21, F = 0, age 31.4 ± 7.2, 20 right-hand dominant, 11 British (white skin), 10 were Asian (brown skin)	controlled, sport (walking, rest, jogging, rest, running, rest), total duration 48 minutes	*r*, BA Plot	HR: during whole protocol right hand device *r* = 0.75 — 0.97, bias range 0 — 11 bpm LoA range [-19 - 7] to [8 40] bpm; left hand device *r* = 0.71 — 0.97, bias range 0 — 12 bpm LoA range [-12 - 6] to [6 42] bpm
(Gorny et al. 2017)	M = 1, Fitbit Charge HR	Polar H6	N = 10, F = 3, age 25.4 ± 3.7	non-controlled, 3 – 6 hours of non-sleeping activities, 1 month continuation in free living conditions without reference	BA Plot	HR: during first 3 - 6 hours Fitbit Charge HR average bias = -5.96 bpm; low HR intensity bias -4.24 bpm LoA -17.4 to 8.95 bpm; moderate HR intensity bias -16.2 bpm LoA -44.0 to 11.6 bpm
(Boudreaux et al. 2018)	M = 7, Apple Watch Series 2, Fitbit Blaze, Fitbit Charge 2, Polar A360, Garmin Vivosmart HR, TomTom Touch, Bose SoundSport Pulse	Ambulatory ECG	N = 50, F = 28, age males 22.00 ± 2.67, age females 22.71 ± 2.99	controlled, sport, 1st part (rest, ergo-cycle with increased load until volatile exhaustion) for 15 minutes, 2nd part (rest, lower and upper limbs resistance exercise with increase load) for approx. 35 minutes	MAPE	HR: MAPE during whole 1st part, 2nd part: Apple Watch Series 2 MAPE = 4.14%, 10.99%; Fitbit Blaze MAPE = 21.06%, 13.74%; Fitbit Charge 2 MAPE = 21.36%, 9.97%; Polar A360 MAPE = 19.48%, 8.66%; Garmin Vivosmart HR MAPE = 25.38%, 10.66%; TomTom Touch MAPE = 12.33%, 19.14%; Bose SoundSport Pulse MAPE = 7.44%, 6.24%

Table 5: Summary of studies evaluating IBI estimation of OHR monitors

Study	**Tested devices (M = number of tested devices, devices' names)**	**Reference device type, name**	**Subjects (N = number of subjects, F = females, age mean ± SD, skin color if available)**	**Protocol description**	**Statistical metrics (according Table 1)**	**Key results (bias [ms], LoA [ms], HR score < 10 bpm [%], MAE [bpm], correct, extra and missing beats [%])**
(Renevey et al. 2013)	M = 1, CSEM proprietary wrist monitor prototype	Holter ECG, Embla Titanium	N = 26, F = N/A, age 48.1 ± 14.4, 22 subjects previously diagnosed with chronic mountain sickness and sleep-disordered breathing.	non-controlled, sleep, totally recorded 13991 minutes of data	ME ± SD of ME, BA Plot	IBI: for all recorded data ME ± SD of ME 0.05 ± 17.96, bias = 0.05 ms LoA -35.7 to 35.81 ms
(Pietilä et al. 2018)	M = 2, PulseOn, Empatica E4	ECG RRI recorder, Firstbeat Bodyguard 2	N = 24, F = 0, age N/A	controlled, daily activity (rest, dish washing, table cleaning, mental arithmetic test, ergo cycle, rest), total duration 53 minutes	MAE, HR score < 10 bpm, correct, extra, missing beats	HR: during whole protocol PulseOn MAE = 1.84 – 4.47%, HR score < 10 bpm = 89.76 – 99.28%; Empatica MAE = 2.50 – 6.16% HR score < 10 bpm = 80.64 – 97.11% IBI: during whole protocol PulseOn beats correct = 76.2 – 90.3%, extra 3.4 – 8.4%, missing = 5.9 – 15.4%; Empatica E4 beats correct = 9.1 – 67.9%, extra 0.3 – 4.2%, missing = 29.5 – 90.6%

4.4 Evaluation of fitness parameters based on heart rate

Bassett et al. (Bassett et al. 2012) have summarized the main issues, recommendations and best practices for the calibration and validation of wearable monitors for fitness parameters, such as VO_{2max} and EE. However, relatively few evaluation studies of consumer OHR wearable monitors have focused on the accuracy of measuring either VO_{2max} or EE using the measured HR as an input parameter. There follows a summary of those studies in which ECG- or PPG-based consumer HR monitoring devices which calculated VO_{2max} or EE were validated against reference measurements.

Crouter et al. (Crouter et al. 2004) evaluated the accuracy of the EE estimation of the Polar S410 chest-strap HR monitor against IC during various exercises using both measured and predicted VO_{2max} and maximum HR (HR_{max}) on a group of the 20 participants (10 male, age 26.0 ± 3.1 and 10 female, age 23.0 ± 2.4). At first, the participants performed standard sports-medicine procedures based on treadmill running until there was volatile fatigue. The initial tests involved lactate level estimation from blood samples and IC based on gas analyses of the subjects' breath. These were used together to estimate the true VO_{2max} and HR_{max} values. In the next stage, the participants performed three submaximal exercises corresponding to the 'moderate', 'hard' and 'very hard' levels of perceived exertion on a treadmill, an ergo-cycle, and a rowing machine, for nine submaximal bouts. During the submaximal testing the subjects' EE was measured simultaneously with IC and with two Polar S410 HR monitors collecting data; one HR monitor was set up to use the predicted VO_{2max} and HR_{max}, and the other was set up to use the measured values. Although there was a strong correlation between the predicted and actual VO_{2max} (r = 0.872, P = 0.001) for the males, there was only a moderate correlation for females (r = 0.477, P > 0.05). In fact, there were no significant (P < 0.05) differences in EE estimation during all the submaximal exercises for males, i.e. in the mean values for both measured VO_{2max} and HR_{max}, and predicted VO_{2max} and HR_{max}. However, the EE estimated for females from the predicted VO_{2max} and HR_{max} values significantly overestimated the mean EE in all three activities, the treadmill (by 2.4 kcal·min^{-1}), the ergo-cycle (by 2.9 kcal·min^{-1}), and the rowing machine (by 1.9 kcal·min^{-1}) (all P < 0.05). The true measured valued of VO_{2max} and HR_{max} for females significantly improved the estimation of mean EE for all exercises, but it still overestimated the mean EE on the treadmill (by kcal·min^{-1}) and the ergo-cycle (by 1.2 kcal·min^{-1}) (P < 0.05).

Montgomery et al. (Montgomery et al. 2009) performed an evaluation of EE and VO_{2max} predictions of the Suunto HR system against a reference gas analysis system during a two-stage incremental running test to establish submaximal and maximal oxygen uptake values for a group of 17 well-trained participants, 10 male, (29.8 ± 4.3 years old) and 7 female (25.6 + 3.6 years old). The EE and VO_{2max} values were calculated at 3 levels by the Suunto system software applying these parameters: basic personal information (BI), BI + measured maximum HR (BI_{hr}), and BI_{hr} + measured VO_{2max} ($BI_{hr + v}$). The results were obtained using linear regression to estimate standard error of the estimate (SEE). The SEE for the VO_{2max} calculations of the Suunto system compared with the true reference values for BI, BI_{hr} and BI_{hr+v} were 2.6, 2.8, and 2.6 kcal·min^{-1} respectively. The SEE of the Suunto system's EE estimation for BI, BI_{hr} and BI_{hr+v} were 6.74, 6.56 and 6.14 kcal respectively. It was concluded that the Suunto system underestimated VO_{2max} and EE by 6 and 13%, respectively. The particular results showed that the validity of EE and VO_{2max} predictions improved with each sequential addition of the measured physiological parameters.

Erdogan et al. (Erdogan et al. 2010) analyzed the EE estimation accuracy of a chest-strap HR monitor, the Polar S810i against IC during submaximal rowing training with 43 over-weight and obese adults (16 male, 27 female) 34.9 ± 5.5 years old. At first, the true VO_{2max} and HR_{max} of the subjects were measured using a standardized VO_{2max} test. The true measured VO_{2max} and HR_{max} parameters were used as the input parameters for HR-based EE estimation and fed into the Polar S810i configuration during another submaximal exercise test. The test consisted of a short warm-up period, followed by 10 minutes rowing at fixed workloads of 50% (low intensity) and after a 20-minute rest, 10 minutes of 70% (moderate intensity) of each subject's predetermined VO_{2max}. The correlation between the mean EE assessments of the Polar S810i and IC at both intensities was strong $r = 0.86$ ($P < 0.001$) and stronger $r = 0.95$ ($P < 0.001$), respectively. The BA Plot analyses of the EE between the chest-strap HR monitor and IC showed a small mean bias (95% LoA) -0.5 (-1.50 to 0.49) kcal·min^{-1} at low exercise intensity and an even smaller mean bias (95% LoA) -0.2 (-1.35 to 0.95) kcal·min^{-1}at moderate intensity.

Robertson et al. (Robertson et al. 2015) examined the accuracy of EE and VO_{2max} assessment using HRV measured with the ECG-based Firstbeat Bodyguard 2 (BG2) device against IC reference measurements during low-intensity walking and a VO_{2max} running test on 24 healthy, non-smoking, 22 ± 2.3-year-old non-elite athletes. Moderate correlation $r = 0.493$, and a small difference in the mean values, 0.61 ml·kg^{-1}·min^{-1} (1.3%), between HRV-based and the actual measured VO_{2max} values were reported during

VO_{2max} test running. HRV-based EE estimation was performed applying true individually measured VO_{2max} and HR_{max} values, and also using age-based HR_{max} and training level-based VO_{2max} parameters. Strong correlations (r = 0.722 at low intensity walking and r = 0.973 at high intensity running) were reported between both the EE estimations, the true measured physiological parameters and the reference method. When user-age based HR_{max} and training level based VO_{2max} parameters were applied, the EE estimation correlation decreased at low intensity, r = 0.598, but at high intensity there was still a strong correlation, r = 0.844.

Rousset et al. (Rousset et al. 2015) compared the TEE assessment of the Actiheart wearable HR monitor against reference measurements in controlled and free-living conditions. The 49 participants spent 17 hours in controlled conditions inside a calorimetric chamber performing pre-defined protocol activities including sleep, rest, office work, eating, walking and running on a treadmill. Furthermore the doubly labeled water (DLW) method was applied in free-living condition for 10-days to trace TEE with a group of 41 participants. In control conditions, an overall MAPE (± SD) 8.6 ± 6.3%, and a BA Plot showing a mean bias (95% LoA) 0.03 (-0.39 to 0.32) $kcal \cdot min^{-1}$ have been reported. In free living conditions, the TEE estimation with the Actiheart wearable HR monitor produced higher errors than those obtained in controlled conditions MAPE (± SD) 12.8 ± 9.1%, a higher mean bias 0.07 and (95% LoA) (-0.69 to 0.54) $kcal \cdot min^{-1}$. Interestingly, a higher TEE MAPE was reported for males than for females in both the controlled (9.7% vs. 7.7%) and free living conditions (15.5% vs 10.0%).

Wallen et al. (Wallen et al. 2016) in their HR estimation accuracy study also evaluated the EE estimation accuracy of 4 wrist-worn OHR monitors, Apple Watch, Fitbit Charge HR, Samsung Gear S, Mio Alpha in a sport protocol. Their reference was IC with a portable gas-analysis system. Table 4 describes the test-subjects' demographics and the protocol activities. Surprisingly, although it was reported that Samsung Gear S does not incorporate HR measurement into its EE estimation, it provided the most accurate average EE estimation at r = 0.86. The correlations between the IC and Apple Watch, Fitbit Charge HR and Mio Alpha were 'very weak' r = 0.16, 'moderate' r = 0.64 and 'weak' ρ = 0.46 respectively. The study suggests that the measurement error in HR and EE estimation may not correlate in all devices.

Hendrikx et al. (Hendrikx et al. 2017) in a clinical evaluation study of the Philips Health Watch wrist-based OHR validated the HR measurement performance and the accuracy of the TEE estimation against IC during a protocol which consisted of various daily and

sport activities. Detailed descriptions of the participants' demographics and protocol activities are shown in Table 4**.** The overall TEE estimation errors for the whole protocol were MAE (± SD) 27.5 ± 28.7 kcal and MAPE (± SD) 10 ± 8.7%.

Besides HR accuracy measurements, Dooley et at. (Dooley et al. 2017) also examined the EE accuracy of three wrist-based consumer OHRs (Apple Watch, Fitbit Charge HR, Garmin Forerunner 225) with 62 participants during a sport protocol, as described in Table 4. Their overall results showed a very wide error range in EE estimation accuracy (MAPE) for all of three devices: Apple Watch 14.07% – 210.84%, Fitbit Charge HR 16.85% – 84.98%, Garmin Forerunner 225 30.77% – 155.05%.

In their consumer evaluation study, Boudreaux et al. (Boudreaux et al. 2018) examined the EE of six consumer OHR monitors with 50 participants during ergo-cycle exercise with increasing loads and lower and upper limbs resistance exercises (see Table 4). The EE estimation error (MAPE) during the ergo-cycle exercises for the Apple Watch Series 2, Fitbit Blaze, Fitbit Charge 2, Polar A360, Garmin Vivosmart HR and TomTom Touch were 21.13%, 72.01%, 75.15%, 38.18%, 63.05% and 41.27%, respectively. Similar EE accuracy among all the tested devices was reported during the resistance exercises. The MAPEs for the Apple Watch Series 2, Fitbit Blaze, Fitbit Charge 2, Polar A360, Garmin Vivosmart HR and TomTom Touch were 42.69%, 49.07%, 47.85%, 52.95%, 57.02%, and 51.54%, respectively.

Most of the studies evaluated the HR-based VO_{2max} and EE parameters in controlled laboratory conditions during VO_{2max} testing or submaximal exercise. This is because specialized equipment is usually required for the reference measurements, as most IC methods are based on the gas analysis of breath. A calorimetry chamber (room) and doubly-labelled water methods were used for reference EE measurements in controlled and free living conditions (Rousset et al. 2015). The accuracy and errors of the VO_{2max} and EE parameters have been reported using various statistical methods and in different units. The most common methods for showing results were the correlation between the tested and the reference devices using a Bland-Altman plot as a graphical representation of the difference between the measurement techniques, including mean bias and LoA, MAE, or MAPE. For chest-strap monitors, a moderate correlation between the IC reference and the HR-based predicted VO_{2max}, was reported for females, r = 0.477, and a strong correlation, 0.872, for males. In low intensity exercise, chest-strap monitor EE assessments in combination with true measured HR_{max} and VO_{2max} correlate strongly with the true measured IC values, r = 0.722 – 0.86. In high intensity exercise there is an even stronger correlation, r = 0.95 – 0.973. For HR-based EE estimation applying age-based

HR_{max} and training index-based $VO_{2max,}$ there are moderate to strong correlations at low and high intensity training, $r = 0.598$, $r = 0.844$ respectively. For the total EE assessment using ECG-based HR during combined sport and daily activities, the bias varies from 0.03 to 0.07 kcal with LoA variation between [-0.69 0.39] and [0.32 0.54] kcal, while the corresponding MAPE is 8.6 – 12.9%. However, a higher error in the EE estimation for males during daily activity protocols has been reported (Rousset et al. 2015) and EE was significantly overestimated during sport activities for females (Crouter et al. 2004). There are relatively few studies evaluating the fitness parameters utilizing wrist-based OHRs. Weak to moderate correlations, $r = 0.16 – 0.64$, have been reported during sport activities, while there is a wide error rate, MAPE 14.07% – 210.84%, between consumer wrist-based OHR monitors which estimate EE based on acquired HR and IC reference measurements. It could be speculated that the larger error levels in the EE assessment MAPE 149.64% – 210.84%, are mainly related to the rest and recovery testing periods of the protocols. Nevertheless, during light to moderate exercise, and vigorous exercise, the lower error ranges were reported to be MAPE 14.07% – 84.98%, and MAPE 19.64% – 30.77% respectively (Dooley et al. 2017). Similar but larger EE estimation error ranges during the ergo-cycle and resistance exercises were reported in another study as being MAPE 21.13% – 75.15%, MAPE 42.69% - 57.02%, respectively (Boudreaux et al. 2018). However, the clinically-validated OHR device combined sport and daily activity testing protocol achieves an EE estimation with a low MAPE 10.0%, (Hendrikx et al. 2017), which accords with the EE assessment of chest strap monitors in similar conditions MAPE 8.6 – 12.9%.

5 Evaluation framework

The Cambridge Dictionary defines evaluation as “the process of judging something’s quality, importance, or value” (Cambridge University Press).

Wearable OHR monitors can be evaluated with various parameters such as their accuracy in measuring the specified variables, their usability, interoperability, wireless connectivity, power consumption, or production costs. This thesis mainly focuses on an objective evaluation of OHR’s accuracy in measuring the HR and IBI variables. The other parameters are not discussed in this thesis.

The international vocabulary of metrology defines accuracy as “the closeness of agreement between a measured quantity value and a true quantity value of a measurand” (Joint Committee for Guides in Metrology 2008). The Oxford Dictionary defines validation as “the action of checking or proving the validity or accuracy of something” (Oxford University Press) while the Cambridge Dictionary defines performance characterization as “how well a person, machine, etc. does a piece of work or an activity” (Cambridge University Press).

Despite the popularity and wide use of wearable OHR sensors nowadays and the importance of their accuracy, there are no commonly agreed guidelines or standards for the systematic evaluation of their accuracy. Several testing, inspection and certification companies offer evaluation, quality-validating services and certification marks for wearable consumer technologies (TÜV SÜD; Bureau Veritas). However, their often comprehensive testing processes usually involve checking different functionalities according to the requirements of international standards, such as battery life-cycle testing, testing the radio-frequency wireless equipment, the specific absorption rate, biocompatibility and mobile application testing. They also perform optional usability tests including user-

friendliness, durability and operating instructions (TÜV SÜD) but they do not objectively evaluate the accuracy of the devices.

The International Organization for Standardization (ISO) describes the requirements for the performance of medical pulse oximeter equipment including a short specification of the pulse rate accuracy (ISO 80601-2-61 2011). The ISO standard 80601-2-61 part 201.12.1.104 defines the reporting of pulse rate accuracy for medical pulse oximeters only as being the root mean squared (RMS) difference of measured paired pulse rate values between pulse oximeter equipment and a reference method (ISO 80601-2-61 2011). However, there are a number of possible reference methods defined in the ISO standard: an electronic pulse simulator, an ECG-based HR, a palpated pulse, a thoracic auscultation, or another pulse oximeter that compares positively with one of the above-mentioned reference methods (ISO 80601-2-61 2011). The American Food and Drug Administration (FDA) also provides guidelines for testing the accuracy of medical pulse oximeters, including pulse-rate measurement (Pulse Oximeters - Premarket Notification Submissions [510(k)s] 2013). The guidelines mainly refer to the above-mentioned ISO standard 80601-2-61, which describes the performance of pulse oximeter equipment. The FDA guidelines recommend testing the accuracy of the pulse rate estimation with a functional tester in a normal set-up. The set-up should represent motion or low perfusion conditions if the equipment has these features. The FDA also recommends following the same ISO-standard accuracy requirements as are defined for assessing the accuracy of blood oxygen saturation (SPO_2) measurement during motion (ISO 80601-2-61 2011) part 201.12.1.102 and low perfusion (ISO 80601-2-61 2011) part 201.12.1.103 in ISO standard (ISO 80601-2-61 2011). However, during SPO_2 measurement in low perfusion and motion conditions, the ISO standard mainly requires that the methods used to establish the accuracy of SPO_2 measurement are described and that an indication of pulsatile signal strength is provided using percentage modulation of the infrared signal. Both the ISO Standard and the FDA guidelines are focused mainly on the measurements of SPO_2 in medical pulse oximeters placed on a finger or an ear. These guidelines for performing pulse-rate accuracy testing are not directly applicable for our purposes, i.e. evaluating the accuracy of the HR estimation of wearable OHR devices. The ISO and FDA guidelines are a good start, but their methodology is created for specific sensor locations (finger and ear), and do not cover the wrist or the forearm, or the selection of the test subjects. In addition, the specifications for the actual test conditions are imprecise or absent.

Bassett et al. (Bassett et al. 2012) have also proposed calibration and validation protocols for wearable sensors focused on monitoring physical activity. However, no specific

requirements for the HR estimation-accuracy of consumer-wearable devices appear to be agreed upon.

A practical, objective framework for the evaluation of the accuracy of optical HR monitors should allow different devices to be quickly and easily compared according to their usefulness and their value for a specific use-case, i.e. they should take into account the user requirements. For example, the following factors should be taken into account:

1. Usage conditions – where will the device be used, e.g. during what kind of physical activity, what will be the temperature, moisture, ambient light, etc.
2. User characteristics – age, gender, skin color, height, weight, BMI, fitness level, wrist size, health status etc.
3. Monitoring duration – is the device meant for long-term or episodic use
4. Accuracy requirements – what level of accuracy is required for the target use-case

In spite of the lack of commonly-agreed guidelines for the evaluation of wearable optical HR monitors, several evaluation studies with varying methodologies have been published recently (section 4.3). In this chapter, an evaluation framework which hopefully will be used for further studies is presented. The framework focuses on ensuring the internal validity of the evaluation (minimizing systematic error “bias”) while also improving its external validity (generalizability of the results to other conditions).

The proposed evaluation framework is divided into three main parts:

1. The design of the evaluation campaign
2. The execution of the evaluation campaign
3. The analysis and reporting of the results of the performance and accuracy of the device

The different parts of the framework are described below and recommendations are derived based on scientific principles, empirical experience and information obtained in earlier evaluation studies specifically focused on the accuracy of optical wearable HR monitoring devices.

5.1 The design of the evaluation campaign

This part of the evaluation framework covers the design of the testing protocol and the selection of test subjects and reference devices.

A key target of the evaluation is its external validity: the extent to which its results are generalizable outside the specific study subjects and protocols. To be useful, a good evaluation study should carefully define the requirements that are relevant to the target use-cases and users, presenting a reasonably wide spectrum of natural variance in different cases while controlling other conditions for repeatability. However, it should be noted that any factors which are controlled, while improving the study's internal validity, may reduce its external validity if they include factors which are not well controlled in the target use-cases (e.g. usage conditions such as temperature, local blood perfusion, etc.). The evaluation campaign should be designed to account for the following factors:

1. Target user profile – age, gender, health status, skin color, fitness level, height, weight, etc.
2. Usage conditions (activities, environment, duration)
3. Reference device

Although consumer-wearable HR monitors are not classed as clinical devices, good clinical practice for investigation of medical devices, and ethical guidelines should be followed when evaluating any devices for human subjects (Ethical principles of research in the humanities and social and behavioural sciences and proposals for ethical review 2009; ISO 14155 2011). The evaluation campaigns are performed with the participation of human subjects and therefore certain legislation and best scientific practices have to be observed. The campaigns should be approved by an agreed-upon competent authority at the local or national level e.g. local ethical review board. Protocols have to be conducted according to the Helsinki declaration (World Medical Association 2013) and subjects have to sign an 'informed consent' form. In addition, the purpose of the study and the expected outcomes have to be clarified, there must be insurance cover, and the responsibilities, risk and benefits of participation have to be clearly defined, as do the procedures for the storage and protection of confidential personal data. These requirements are typically defined at the national level (Ethical principles of research in the humanities and social and behavioural sciences and proposals for ethical review 2009; Clinical trial information leaflet and consent 2016).

5.1.1 Design of the testing protocol

The testing protocol must define the activities and the conditions under which the evaluation is carried out. It must match the requirements mentioned above and fulfill the following criteria:

1. Standardization
   Refers to which activities are defined to be performed, how accurately, and their duration.
2. Repeatability
   Refers to the reproducibility of the target use-case scenario and the defined activities and conditions.
3. Representativeness
   Refers to definition of usage conditions and possible impact of external factors.

Controlled or non-controlled protocol execution

Many of the trials used controlled protocols with continuous supervision, typically specified indoor gym or lab activities. Although controlled conditions are more standardized, repeatable and well-annotated, the external validity of controlled protocols is more difficult to generalize for common usage conditions and typical user behavior, e.g. supervision of the test subjects during execution of the test reduces the number of uncontrolled hand movements in order to decrease the level of possible artifacts and measurement errors. Non-controlled protocols, typically everyday life, sleep or outdoor sports activities, without continuous supervision and without any detailed specification of the activities is closer to the actual target use-case conditions. However, although the external validity of such protocols is higher as they are more representative of the target use-case conditions, they are weaker in terms of repeatability and standardization. Any interpretation of the results of non-controlled protocols relies on user-subjective annotations of possible unexpected events during the test.

External environment factors that can affect the signal quality and the accuracy of the HR estimation, such as the outdoor temperature or ambient light interference, also need to be specified and controlled as closely as possible.

Testing scenario specification

The activities in the testing protocol must directly reflect, or at least closely simulate, real usage conditions. A typical testing protocol, which describes a protocol derived from

sport physiology testing guidelines (Winter et al. 2006; Nieman 2011), is presented in Table 6. The main advantages of indoor testing are the fact that it is relatively simple to standardize the test conditions (e.g. temperature and ambient light), the researcher has full control of how the test is executed, and the tests can easily be repeated. However, such protocols control several factors which cannot be controlled during typical real-user scenarios (e.g. temperature variations and related variations in skin blood perfusion, varying ambient light level, clothing, activity pattern such as step rate and variations in style when exercising, e.g. on a treadmill or an outdoor path, etc.). Although outdoor testing better represents real user conditions (outdoor jogging, cross-country skiing, inline-skating) many external factors may not be controlled and these may vary widely, so such evaluation studies require much larger samples and it is more challenging to define the practical limitations of the test conditions. Therefore, such evaluations are scarcer. Daily activities represent long-term scenarios aimed at revealing any potential inaccuracy in the device being tested during random unspecified wrist movements (e.g. washing dishes or typing on a keyboard). These scenarios can be performed within defined protocols in specialized laboratories; unspecified hand movements can be simulated with a Rubik's cube game, for example. Motionless (e.g. sleep or awake rest) protocols are used for testing beat-to-beat detection accuracy and the long-term resting HR level with an unobtrusively-worn test device.

It is important to include two special activities at the beginning of the protocol: a warm-up activity and a synchronization activity. The former aims to standardize the subject's blood perfusion and body temperature, while the latter (e.g. several squats) will help synchronize signals between the tested and reference devices during any subsequent signal processing.

The activities in the protocol should be performed in a predefined order to ensure that the measurements are standardized and repeatabable. Some activities, typically exercises such as walking or ergo-cycling, might improve blood perfusion, so subsequent test activities must take account of this in order to achieve more accurate HR estimation. The results estimated according to the example protocol proposed in Table 6 might be biased towards a lower measurement error. This is caused by the correlation of human HR and running speed while systematically increasing the running speed in the protocol. Moreover, no breaks between running speeds are included; these are typical for interval training which represents the most difficult testing scenario.

Table 6: Example of testing protocol for indoor testing scenario

Activity	**Duration**
Synchronization – 5 squats and rest	1 min
Warm-up exercise – 6 km·h^{-1} running	5 min
Still standing	1 min
Walking on a treadmill at 3 km·h^{-1}, 0% inclination	3 min
Walking on a treadmill at 3 km·h^{-1}, 10% inclination	3 min
Walking on a treadmill at 5 km·h^{-1}, 0% inclination	3 min
Walking on a treadmill at 5 km·h^{-1}, 10% inclination	3 min
Running on a treadmill at 9 km·h^{-1}, 0% inclination	3 min
Running on a treadmill at 11 km·h^{-1}, 0% inclination	3 min
Rest sitting	4 min
Cycling 60 rpm, 75 Watts resistance	3 min
Cycling 90 rpm, 75 Watts resistance	3 min
Rest sitting	4 min

The order in which the activities are performed in the protocol may affect the results of the evaluation. This variation is hard to control unless the order of the activities is randomized. However, randomizing the order of the activities can cause problems for the practical execution of the evaluation campaign. The order of the activities in the protocol needs to be taken into account in order to reduce the output error, but this can systematically and deliberately bias the results. For example, activities designed for testing the sensitivity of hand movements' artifacts performed after a warm-up exercise might increase accuracy of the HR estimation, but decrease the external validity of the protocol.

5.1.2 Selection of the test subjects

Non-probabilistic techniques such as convenience, purposive or quota sampling (Martinez-Mesa et al. 2014) are usually applied when selecting the test subjects. The selected test subjects must represent the target-user population in the following key aspects.

1. Actual health status and fitness level
2. Age
3. Gender
4. Skin color classified according to a defined scale, e.g. the Fitzpatrick scale (Fitzpatrick 1988)
5. Wrist circumference and its anatomical shape (bony, normal, fatty)

Particular purposive sampling methods such as maximum variation and homogeneous or typical-case sampling can be useful to fulfill specific population requirements (Etikan et al. 2016). Nevertheless, it must be remembered that it is harder to generalize the results of tests using non-probabilistic sampling methods. The minimal required number of test subjects depends on the primary aim of this kind of evaluation study, which is to provide descriptive statistical outcomes. However, the output error metrics are quantitative variables. It is assumed that evaluation studies of wearable OHR monitors are typically based on a mixed research design that is based on both quantitative and qualitative research approaches. A typical sample size depends on the protocol, but a low range of 10 to 30 subjects is required to obtain basic evidence and statistics in controlled conditions. However, for non-controlled conditions, or for randomization, much larger samples are required. The power analyses for calculating the minimal number of subjects are not applicable unless there is sufficient information about the target population. The basic empirical research rules for a group of test subjects are to maintain an equal gender balance and group the subjects by a maximum ± 10 years deviation from the mean age of the sample. In addition, if the aim of the research is to test some special characteristic e.g. different skin types, the characteristic should be represented with at least 5 subjects per specific property.

Nearly all measurements are influenced by different sources of variation: biological assays include pre-analytical variation, analytical imprecision, analytical bias, within-subject normal biological variation and between-subject variation (Fraser 2001). Analytical imprecision can be reduced by averaging replicates (e.g. repeating measurements) from the same sample (Monach 2012). The effects of both within-subject biological variation and analytical imprecision can also be alleviated by averaging measures performed repeatedly over time (Monach 2012).

5.1.3 The tested devices

The tested device has to be used under the same conditions (temperature, moisture) and for the same purpose (exercise or daily monitoring) for which it was originally designed according to the user manual. If the device is not operated according to the manufacturer’s instructions, the HR measurement might be inaccurate and this can bias the statistics. The device settings for age, gender etc. have to be correct. The device has to be attached and worn as defined by the manufacturer, and its position and the tightness of the strap(s) needs to be properly checked if the test is to be repeatable. The measured variables and their output time resolution should be clearly defined for further statistical analyses. It is preferable that the data is automatically gathered from the tested device

rather than marking down numbers manually from the device's display, which might produce various inaccuracies.

5.1.4 Reference devices

The ECG-based reference devices classified as "gold standard" are mandatory: medical or scientific ECG recorder, chest-strapped HR monitors or disposable-electrode IBIs recorders. Any potential inaccuracy in the reference device includes an additive error in the final measurement error of the tested device. Prior evaluation of the ECG-based chest-strap or disposable-electrode wearable HR monitors against certified medical ECG recorders is recommended as this will reveal any unknown inaccuracies in the reference device's measurements. Several scientifically-evaluated ECG-based wearable HR monitors validated against "gold standard" medical ECG signal recorders are currently available (Kingsley et al. 2005; Vanderlei et al. 2008; Parak & Korhonen).

5.2 The execution of the evaluation campaign

An evaluation campaign should follow good clinical practice for the testing of a medical device for human subjects (ISO 14155 2011) in order to preserve the safety of the human participants. Operators who conduct controlled protocols have to clearly understand how to execute the protocol procedures and the functionality of the tested and reference devices. During non-controlled evaluation protocols it is important to provide all of this information to the test subjects in advance, especially anything that can help them deal with potential problems when doing the test. Any information which deviates from the defined protocol must be collected by the operators or the test subjects and reported for further analysis and data interpretation. The organizer of an evaluation campaign is obliged to supervise and check all the initial conditions, and they should also perform random or regular visits to control the evaluation process and avoid any potential errors (Delgado-Gonzalo et al. 2018).

5.3 Pre-processing the measured signal for evaluation

5.3.1 Time synchronization

The measurements of both the tested and reference devices need to be precisely synchronized in order to calculate the statistical error metrics. Improperly synchronized data can bias the results.

Time synchronization between one or several tested devices and a reference device can be challenging for HR, especially for beat-to-beat evaluation. Internal device clocks are not usually exactly synchronized to each other. There are two main problems with time synchronization:

1. Bias – difference in start time, which is a common problem
2. Drift – difference in clock rate, usually unimportant for short-term measurements but may be relevant for long-term monitoring

The estimation of the time lag between the measured signals solves the bias problem. The time lag can be estimated by computing a cross-correlation of the signals or by applying minimal error matching between signals. In both methods, all of the acquired signals are resampled at a higher sampling frequency to achieve a precise estimation of the time lag. If the devices also provide simultaneous sampling of motion signals (i.e. acceleration), the sensor signals should be used within a synchronized activity period of the protocol to estimate the time lag (Bannach et al. 2009).

The clock drift problem is handled by splitting the measured signals into smaller time periods. Then the time lag is estimated for each segment separately. Further statistics are also calculated within the synchronized segments and finally summarized for the overall results.

Both the reference and tested signals have to be re-sampled at the same regular sampling frequency in order to be able to perform a statistical evaluation of the measurement accuracy. If the nominal sampling rate of any device differs from a declared value, it should be corrected by a scaling factor which based on the difference between the nominal and declared values.

5.3.2 Reference signal processing

Processing the reference device signals requires that the beats are detected in RAW ECG signals to estimate the corresponding IBIs. Manual checks are performed to guard against any potential errors with the automatic methods. Kubios HRV, Matlab-based HRV analysis software provides suitable methods and a user interface for these purposes (Tarvainen et al. 2014).

Further IBIs are cleaned up with an appropriate artifact-correction algorithm (Saalasti 2003; Saalasti et al. 2004). Ectopic beats representing physiological arrhythmia also have to be removed from the reference signals with an appropriate algorithm (Mateo & Laguna 2003). Then, instantaneous reference HR values are estimated from IBIs, and these are further averaged in specific time-windows.

The instantaneous or time-averaged HR values may also be directly extracted from designed services or the proprietary SW of the reference device, especially for chest-strap or disposable-electrodes HR recorders.

5.4 Evaluating accuracy

The accuracy of a tested device can be calculated and represented using the various statistical metrics and methods defined in section 4.1 of this thesis

Firstly, it is important to establish a specific time-error range in order to use HR error statistical metrics. The recorded data from both the tested and reference devices are usually split into short (typically 5 s) time-windows with no overlap. The average HR value is estimated within these short time-windows, and these window-averaged HR values are used as inputs for statistical methods for estimating error metrics.

The most suitable error metrics for characterizing the accuracy of HR estimation are the mean absolute error (MAE), relative or absolute HR scores (named also HR reliability), or HR accuracy as a complement of mean absolute error (MAPE). The MAE is a useful generic error metric that describes the average difference between the true and estimated HR, and shows the first norm of the average disagreement between devices. The absolute HR score < *X* bpm or relative HR score < *X* % error is a practical metric which describes how often the estimated HR value is as close to the real HR value as is required for the target use-case. The error acceptance threshold, *X*, should be selected

based on the target use-case. For consumer HR monitoring, 10 bpm can be considered a suitable target level for recreational wellness applications. For more demanding applications, some other levels, e.g. 5% may also be considered. The HR accuracy defined as a complement of the relative error (MAPE) can express an appropriate performance level for the users of consumer devices. The estimation of IBI accuracy primarily requires that the statistics for missing, correct and extra detected beats are calculated. MAE, MAPE or a mean ± standard deviation between pairs of IBIs measured by the tested and the reference devices are suitable error metrics for describing the accuracy of IBI estimation. Beat detection algorithms do not usually incorporate an artifact-correction method for detected beats. It is therefore also possible to determine the efficiency of the artifact corrections by comparing the statistical error metrics related to IBI estimation accuracy before and after applying beat-detection artifact-correction methods. The standard for pulse rate measurement in clinical devices requires that the RMSSD of HR pairs is calculated (ISO 80601-2-61 2011). RMSSD measures the standard deviation of the prediction errors and is disproportionately affected by larger errors.

The BA-plot is also an appropriate analytical tool for visually showing agreement between pairs of tested and reference HR or IBI values, including bias and 95% limits of agreement. This method has been widely used and is highly recommended for assessing the measurement error in sport medicine (Atkinson & Nevill 1998). However, using a BA-plot as a method for comparison of different measurement techniques should be considered carefully because it is not always suitable for all applications, such as an evaluation of the cross-validation of regression models (O'Connor et al. 2011), comparison of measurements techniques with different units (Hopkins et al. 2009), or for using a BA plot-based bias and confidence interval to calibrate the devices (Hopkins et al. 2009; Ludbrook 2010; Ludbrook 2010). The method can easily be misused and lead to incorrect conclusions, especially because of the bias involved in estimating the BA limit of agreements (Ludbrook 2010; O'Connor et al. 2011). Linear regression analysis should be used for re-calibrating devices because: it eliminates the impact of substantial random error (Hopkins 2004), it better handles problem with proportional bias between measurements (Hopkins et al. 2009; Ludbrook 2010), and the regression validity analyses can be combined with published validity regression statistics for the inaccurate measurements in order to correctly estimate the validity regression statistics for the new proposed measurement technique (Hopkins et al. 2009). Ludbrook (Ludbrook 2010) has proposed instructions on how to properly construct bias and 95% limits of agreement lines in a BA-Plot according to the type of application and the distribution of the analyzed data, especially with regard to the fixed and proportional bias between measurement techniques.

6 Summary of publications

6.1 Evaluation of HR, EE and VO_{2max} during sports

Publications I, II and V evaluated the HR-estimation accuracy of consumer OHR monitors against gold standard ECG-based reference HR measurements. Publication V examined the accuracy of the EE estimation error using OHR, as well as VO_{2max} estimation error based on the OHR and GPS speed estimated with a mobile phone application during submaximal exercise. Four different consumer OHR monitors were evaluated in these sport-based studies.

6.1.1 Evaluation methodology

Table 7 contains summarized descriptions of all the evaluated OHR devices (Figure 10). The descriptions of the reference ECG-based devices (Figure 11) are presented in Table 8**.** In Publication V, the reference EE was measured with the IC system Metalyzer 3B, Metasoft Studio 4.8, Cortex Biophysik**,** Germany. The blood sample analyses were performed with Biosen C_Line, EKF Diagnostic, Germany, and the outdoor global positioning system- (GPS)-based speed was tracked by a reference device Polar V800. The speed input values for the VO_{2max} estimation algorithm were calculated in Android QT mobility library (The Qt Company) from the GPS location in a Samsung S3 Mini Galaxy mobile phone.

The evaluation protocols in Publications I and II consisted of controlled indoor laboratory activities. The protocol for Publication I also included the simulation of hand movements during a Rubik's Cube game and the impact of lying in bed in different positions. Publication II also evaluated the performance of devices in different non-controlled outdoor testing conditions during various sports. The test protocol for Publication V included con-

trolled VO_{2max} and HR_{max} maximal testing in sport laboratory conditions and semi-controlled outdoor condition to evaluate the accuracy of VO_{2max} estimation during submaximal self-paced running in real winter conditions

Table 7: Properties of the tested devices

	Mio Alpha	**Scosche myRhythm**	**Mio Link**	**PulseOn**
Source light and #LEDs	2 green LEDs	2 IR LEDs	2 green LEDs	2 green and 1 IR LEDs
Number of PDs	1	1	1	1
Location	Wrist	Forearm	Wrist	Wrist
Band type	Rubber silicon	Textile	Rubber silicon	Textile
Wireless connectivity	Bluetooth / ANT+	Bluetooth	Bluetooth / ANT+	Bluetooth
Data storing	Real time streaming	Real time streaming	Real time streaming	Internal memory buffer
Display	Dot - Matrix LCD	No	No	OLED
Size (w x l x h) [mm]	44 x 42 x 17	55 x 49 x 13	25 x 46 x 10	29 x 32 x 12
Weight [g]	56g	29g	33g	29g

Table 8: Properties of the reference devices

	Embla Titanium	**Firstbeat Bodyguard 2**	**Polar RS800CX**	**Garmin Forerunner 610**
Device type	Portable laboratory signal recorder	Chest based disposable electrodes HR recorder	Chest strap	Chest strap
Measured signals	2 channel ECG	RRI, RAW acceleration	HR / RRI	HR / RRI

Figure 10: Evaluated OHR devices

Firstbeat Bodyguard 2
(Firstbeat Technologies Ltd.)

Polar RS800CX
(Polar Electro Oy)

Garmin Forruunner 610
(Garmin Ltd.)

Figure 11: Reference ECG-based devices

Table 9 summarizes the information about the tested devices, the reference devices and the characteristics of the participants for the testing protocols in Publications I, II and V. The participants were healthy, relatively young or middle-aged non-smoking adults without any health issues. Most of the participants were students and employees at Tampere University of Technology, PulseOn and local sports clubs. The testing subjects were Caucasian with Fitzpatrick skin scale classifications of 1 – 3. The experimental procedures performed in Publications I, II and V complied with the principles of Helsinki Declaration of 1975, as revised in 2013. All subjects gave informed consent to participate and they had a right to withdraw from the study at any time. Their information was anonymized prior the analysis.

Table 9: Summary of HR performance testing campaigns (Publication I, II and V)

Pub.	Tested devices	References device	Testing subjects	Protocol description
I	Mio Alpha Scosche myRhythm	Embla Titanium	N = 21, F = 6, age 31.3 ± 10.7	controlled, sport (rest, walking, jogging, running, Rubik's cube play, ergo-cycle), duration 50 minutes
II	PulseOn Mio Link	Polar RS800CX	N = 19, F = 10, age 28.30 ± 5.6	controlled, sport (rest, walking, jogging, running, ergo-cycle), duration 39 minutes
	PulseOn	Polar RS800CX Firstbeat Bodyguard 2 Garmin Forerunner 610	N = 8, F = 2, age 30.9 ± 10.7	non-controlled, sport (outdoor walking, running, cycling), totally 24 events
V	PulseOn	Polar RS800CX	N = 24, F= 11, age 36.2 ± 8.2	controlled, sport (warm-up jogging, rest, VO_{2max} running with increasing speed by 1 $km \cdot h^{-1}$ until volatile fatigue), duration 14 and then 3 minutes for each speed level

For further comparison and statistical calculation the analyzed signals were resampled at 10 Hz sampling frequency, averaged and smoothed out by moving the average filter in a five-second window. Synchronization between the tested and referenced signals was performed by applying a cross-correlation function or minimal error matching between signals. In Publication I, the reference ECG signal analyses were performed by the Kubios HRV tool (Tarvainen et al. 2014). The R-Peaks for HR calculation were detected with a built-in QRS detection algorithm (Pan & Tompkins 1985). All the R-peaks in the reference signals were verified manually. Physiological arrhythmias (ectopic beats) detected by the heart-timing signal algorithm (Mateo & Laguna 2003) were excluded from further calculation of the statistical metrics. In Publications II and V, the RR intervals recoded with reference devices were corrected by applying an RR-interval artifact-correction algorithm based on neural networks and a physiological model (Saalasti et al. 2004) implemented in Firstbeat Sports SW, Firstbeat, Jyväskylä Finland.

All of the HR results statistics were calculated in a 5s average HR window without overlaps. Reliability and accuracy were used as common metrics for comparing the performances of the tested OHR monitors between themselves, and for comparing the performances of the tested devices and the designated ECG-based reference. In addition, in Publication I other metrics such as ME, MAE, MPE and MAPE were estimated. Individual results are provided for the different activities in the protocol and the overall results are over the whole protocol. In order to make the results of the non-controlled outdoor tests in Publication II easier to report and compare, the activities were divided into three main groups (running, biking, walking) according to the dominant activity during the measurement period.

The accuracy of VO_{2max} assessment was verified using age-based HR_{max} (Tanaka et al. 2001) and the measured HR_{max} as input parameters for the computing algorithm. The EE estimation was evaluated during light-intensity (before aerobic threshold) and medium- to heavy-intensity (between aerobic and anaerobic threshold) exercise, as is recommended for physiological exercise testing (Jeukendrup & Wallis 2005; Nummela 2007).

The bias, standard deviation, MAE and MAPE were calculated to describe the error between the EE and VO_{2max} estimated from the OHR device using physiological modelling and reference measurements from gas analyses. In addition to error calculation, the statistical significance of the different measurement methods was verified using parametric and non-parametric tests based on the distribution of a particular dataset. The normality of the distributions has been tested with the Shapiro-Wilk testing method. The Spearman and Pearson correlation coefficients were utilized to examine the agreement between measuring parameters from the OHR and the standard reference method. All the statistical tests were two-sided and the significance level was set at $p < 0.05$. In addition, BA-plots were used for visual inspection of the agreement between the measurement methods and datasets.

6.1.2 Summary of results (HR vs different devices, EE, VO_{2max})

Table 10 and Table 11 presents a summary of the main statistical parameters for describing the accuracy of HR estimation of the devices evaluated in Publications I, II and V during different protocol activities.

Table 10: Key results of the HR performance of the tested devices (Publication I)

			Scosche myRhytm		Mio Alpha	
Pub.		**Activity**	**HR Score < 10% bpm [%]**	**Accuracy [%]**	**HR Score < 10% bpm [%]**	**Accuracy [%]**
I	**Indoor**	**Rest**	83.9	94.0	84.9	94.7
		Walking	81.8	89.5	87.2	94.4
		Running	93.3	96.2	96.2	97.6
		Cycling	97.4	98.3	91.7	94.5
		Rubik's cube	91.8	96.1	72.3	91.6
		Entire Protocol	86.3	93.2	87.5	94.8

Table 11: Key results of the HR performance of the tested devices (Publications II and V)

			PulseOn		Mio Link	
Pub.		**Activity**	**Reliability [%]**	**Accuracy [%]**	**Reliability [%]**	**Accuracy [%]**
II	**Indoor**	**Rest**	97.9	97.1	97.4	97.3
		Walking	90.8	95.8	73.7	90.2
		Running	99.4	98.0	99.8	98.8
		Cycling	96.0	96.8	97.0	97.7
		Entire Protocol	94.5	96.6	86.6	94.3
	Outdoor	**Walking**	94.1	96.6	N/A	N/A
		Running	99.1	97.9	N/A	N/A
		Cycling	95.2	97.3	N/A	N/A
		Mean	97.8	97.6	N/A	N/A
V	**VO_{2max} lab**	**Rest when standing**	96.9	97.1	N/A	N/A
		Ramp-up running	95.3	98.3	N/A	N/A
		Entire protocol	95.4	98.1	N/A	N/A

Publication I evaluated the Mio Alpha wrist-based and Scoshe myRhythm forearm-based OHR monitors. Both devices achieved satisfactory overall performance at 87 – 88%, within <10% of the score of the true reference. The wrist-based device performed better during walking and running activities, while the forearm-based device was more accurate during cycling and the hand movement exercises, such as when playing with the Rubik's cube.

Publication II evaluated the PulseOn and Mio Link wrist-based OHR monitors. The overall reliability parameters for the PulseOn and Mio Link were 94.5% and 86.6% respectively. Both devices were less reliable for walking (PulseOn 90.8%, Mio Link 73.7%). This could be due to lower perfusion because of the order of activities in the protocol. The non-controlled outdoor testing results showed high reliability > 95% for the PulseOn device during walking, running and cycling activities. In the study presented in Publication V, the PulseOn device accurately measured HR during maximal ramp-up running exercises. The reported results were 95.4% for reliability and 98.1%, for accuracy.

The overall HR estimation accuracy for all four devices over the entire protocols was in the range of 93.2% to 98.1%, corresponding to a MAPE of 6.8% – 1.9%.

The reported EE MAPEs during light-intensity and medium- to heavy-intensity exercise were 16.5% and 6.7%, respectively. The MAPEs of the VO_{2max} estimated during submaximal exercise compared to the reference measurement using the GPS speed combined with individually measured HR_{max} and age-based HR_{max} were 5.2% and 5.9%, respectively. The submaximal exercises were performed in outdoor winter conditions. Moderate to strong correlations with the reference were observed for both the EE and VO_{2max} parameters. Table 12 presents the MAPE and the correlation with IC standard reference measurements for EE estimation at both exercise intensities, and for the VO_{2max} estimations including both the true measured and age-based HR_{max} input parameters.

Table 12: VO_{2max} and EE estimation accuracy key results (Publication V)

VO_{2max} estimation				
		All	**Male**	**Female**
Age-based HR_{max}	**MAPE [%]**	5.9	5.2	6.8
	Correlation [-1, 1]	ρ = 0.87	*r* = 0.73	*r* = 0.63
Measured HR_{max}	**MAPE [%]**	5.2	4.7	5.8
	Correlation [-1, 1]	ρ = 0.86	*r* = 0.77	*r* = 0.69
Energy expenditure estimation				
		All	**Male**	**Female**
Light intensity	**MAPE [%]**	13.05	15.28	10.65
	Correlation [-1, 1]	ρ = 0.77	*r* = 0.88	*r* = 0.79
Heavy intensity	**MAPE [%]**	6.7	8.2	5.1
	Correlation [-1, 1]	*r* = 0.97	*r* = 0.93	*r* = 0.99

6.1.3 Conclusions

It can be concluded that optical wearable HR monitors are suitable for monitoring HR during endurance-sport exercises. However, their accuracy might significantly decrease during non-rhythmic exercises or voluntary hand movements. It is also suggested that the optically monitored HR is sufficiently accurate to reliably estimate EE and VO_{2max} if analyzed with an appropriate combination of algorithms. The VO_{2max} during submaximal running exercise can be reliably assessed using OHR and a mobile phone's GPS speed measurements.

6.2 Evaluation of beat-to-beat detection accuracy during sleep (Pub III)

Publication III evaluated the accuracy of the beat-to-beat and HRV parameters of the PulseOn OHR monitor against ECG-based reference measurements during normal sleep conditions. The sleep quality parameters, such as relaxation time, stress time, training effect and recovery index (Firstbeat Technologies Ltd.) were also calculated using PPG and ECG beat detection methods and the results were compared to each other.

The Firstbeat Bodyguard 2 device (described in Table 8**)** handled the reference measurement of the RR interval series. Ten healthy volunteers (8 male and 2 female, 35 ± 10.3 years old) participated in the study as test subjects. In total, 13 recordings were performed and the average non-stop recorded sleep time for all the subjects was 5.1 ± 1.2 hours. The recordings were performed in non-controlled conditions in the subjects' normal bedrooms in their homes. The experimental procedures performed in Publication III complied with the principles of Helsinki Declaration of 1975, as revised in 2013. All subjects gave informed consent to participate and they had a right to withdraw from the study at any time. Their information was anonymized prior the analysis.

The artifact-correction method (Saalasti et al. 2004) and ectopic beats exclusion algorithm (Mateo & Laguna 2003) were applied to both the PulseOn and reference device IBI datasets. An eventual timed drift between the PulseOn and the reference device was compensated for by splitting the data into 5-minute intervals. Minimizing the mean difference between the detected beats was used to synchronise each of these intervals separately.

The proportions of correct, missing and extra detected beats were calculated for the processed and synchronized signals. The number of missing and extra beats against the reference measurements were determined with the methodology described in detail in section 4.1**.** Moreover, ME, MAE, MPE, MAPE error metrics and a comparison of the RMSSD parameters were determined for the corresponding beat intervals.

Table 13: Beat-to-beat detection accuracy of the PulseOn device

	Beat–to–beat detection	
	Before artifact correction	**After artifact correction**
Correct beats [%]	99.42	99.57
Extra beats [%]	1.93	0.72
Missing beats [%]	0.58	0.43

The results in Table 13 show that even without the artifact-correction algorithm, the PulseOn device detected 99.42% of the heartbeats correctly in comparison with the reference device. After artifact correction of both signals the accuracy increased slightly to 99.57%, there was a relative decrease of false positive beats from 1.93% to 0.72%, and the number of false negative fell from 0.58% to 0.43%.

Table 14 presents a statistical comparison of accuracy between pairs of synchronous IBIs. The five-minute intervals containing only the correctly detected beats by the PulseOn device, with one corresponding reference beat, were included in the statistics. The overall mean errors ± SD were -0.32 ± 14.40 ms before and -0.33 ± 11.74 ms after artifact correction. The results are also shown in the BA Plot (Figure 12), which includes the distributions of errors and IBI durations.

Table 14: Beat-to-beat interval estimation

	Beat–to–beat interval estimation	
	Before artifact correction	**After artifact correction**
ME [ms]	-0.32	-0.33
Error SD [ms]	14.40	11.74
MAE [ms]	6.68	5.94
MPE [%]	-0.03	-0.03
MAPE [%]	0.62	0.56

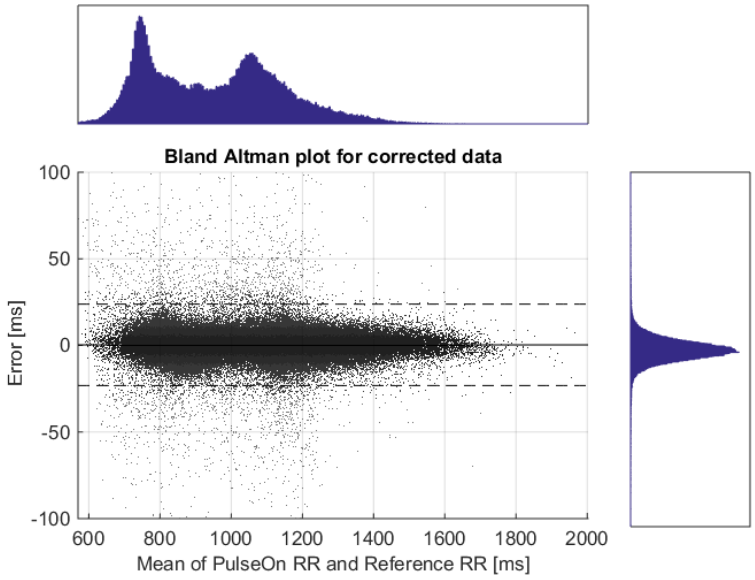


Figure 12: Bland - Altman plot comparing the reference ECG-obtained RRI to the PPG-obtained IBI, for artifact corrected data. Confidence interval ($\mu \pm 2\sigma$) depicted by the dashed lines) is [−23.15, 23.83] ms (Parak et al. 2015). © 2015 IEEE. Reprinted with permision.

The RMSSD estimation, which is one of the HRV measures, was compared between the detected PPG- and ECG-based IBIs. Very small differences were observed between the RMSSD parameter estimated with the PPG-based PulseOn device and the ECG-based reference, 4.2 ms for artifact-uncorrected and 3.1 ms for artifact-corrected data respectively.

Table 15 compares the overall sleep quality parameters determined with Firstbeat Sports SW, Firstbeat, Jyväskylä Finland for ECG- and PPG-based IBI series. Since the recordings were done at night, the results correspond to the amount of relaxation and the duration of the sleep with a low training effect and low average HR.

Table 15: Overall sleep parameter comparison between the PulseOn and Reference devices

	Sleep parameters	
	PulseOn device	**ECG Reference**
Relaxation time [min]	195.38	196.31
Stress time [min]	74.53	82.53
Scaled Recovery index (%)	100	100

It may be concluded that the newer PPG-based OHR monitors are suitable not only for HR measurement during exercise, but also for monitoring HRV accurately and reliably during sleep and no motion. The results show very good accuracy and only a small error in beat detection and IBI estimation when compared to standard ECG-based reference RRI measurements.

6.3 Power saving for monitoring daily life (Pub IV)

Publication IV presented an algorithmic approach to semi-continuous HR monitoring with the aim of reducing the power consumption during long-term OHR monitoring. An experimental evaluation of the semi-continuous HR algorithm ("sampled HR") was performed on various datasets containing data from different activities.

The sampled HR algorithm has been evaluated against continuous OHR estimation and an ECG-based reference on various datasets covering a wide range of activities. It was expected that the designed algorithm with the aim of a faster convergence time would not perform as well as the continuous algorithm. The loss in accuracy was made up for by the reduction in power requirements during the HR estimation and the sampling of the PPG signal. All of the following analyses and experiments were performed offline. A continuous HR was derived from the PPG signal using PulseOn's PPG algorithm. The sampled HR was estimated from the PPG signal using the algorithm presented in the publication.

Figure 13 shows the Sampled HR estimation process in two sampling rounds. The sampling and estimation intervals start at zero seconds and 60 seconds. In the first round, a reliable HR was found after 9 seconds. In the second round, no reliable HR was found and therefore the algorithm output was the value after 20 seconds of the sampling period. The 60-second semi-continuous sampling interval and 20-second timeout were also used in offline experiments.

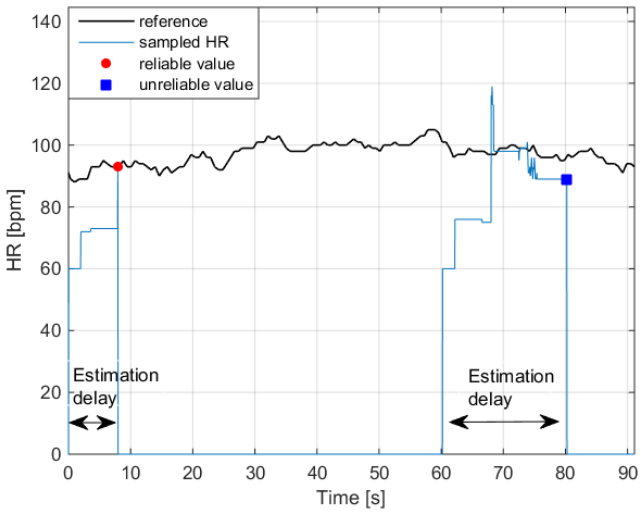


Figure 13: Heart-rate estimation example for the proposed algorithm (Tarniceriu et al. 2016). © 2016 IEEE. Reprinted with permision.

The OHR device used in this study was the PulseOn. Table 16 presents a summary of the four main testing datasets including a description of the activity, the subjects' demographics, the reference device (details described in Table 8) and the duration of recorded signals. The dataset groups were created according to the main corresponding activities. The experimental procedures performed in Publication IV complied with the principles of Helsinki Declaration of 1975, as revised in 2013. All subjects gave informed consent to participate and they had a right to withdraw from the study at any time. Their information was anonymized prior the analysis.

Table 16: Dataset description used for evaluation of "Sampled HR" Algorithm

	Datasets descriptions		
Dataset name	**Protocol description**	**Subjects**	**Reference device**
Indoor lab.	controlled, sport (rest, walking, jogging, running, ergo-cycle), duration 39 minutes	N = 19, F = 10, age 28.3 ± 5.7	Polar RS800CX
Outdoor	non-controlled, sport (outdoor walking, running, cycling), totally 28 events	N = 9, F = 1, age 33.5 ± 10.3	Polar RS800CX Firstbeat Bodyguard 2 Garmin Forerunner 610
Sleep	non-controlled, sleep at home, average time 5.1 ± 1.2 hours per subject, and the total duration 66.5 hours	N = 10, F = 2, age 35.9 ± 10.3	Polar RS800CX Firstbeat Bodyguard 2
Daily activity	non-controlled, daily activity, total duration 17 hours	N = 3, F = N/A, age N/A	Polar RS800CX Firstbeat Bodyguard 2

The HR estimation accuracy was reported in the statistical metrics MAE, reliability and accuracy. In addition, the following values for the sampled HR were computed:

- Reliable estimation delay: the average time duration required to obtain a reliable HR value.
- Reliability rate: the percentage of 60-second intervals for which a reliable HR was found

Table 17: Performance comparison between continuous and sampled-mode HR estimation

Dataset	Mode	MAE [bpm]	Reliability [%]	Accuracy [%]	Reliable estimation delay [s]	Reliability rate [%]
Indoor lab.	Continuous	3.4	93.9	97.1	N/A	N/A
	Sampled	4.2	92.4	96.3	7.8	89.6
Outdoor	Continuous	3.1	92.8	97.4	N/A	N/A
	Sampled	4.5	89.1	96.6	7.1	89.8
Sleep	Continuous	1.8	98.5	96.6	N/A	N/A
	Sampled	1.3	99..5	97.6	9.3	94.4
Daily activity	Continuous	3.3	93.2	95.4	N/A	N/A
	Sampled	2.9	94.7	95.9	10.0	77.1

The performance results (Table 17) of the sampled HR show slightly higher MAE, and lower reliability, during sports activities. However, they show a lower MAE and higher reliability during common everyday activities and sleep. The estimation delay of the sampled HR was in the range of 7.1 seconds (sport) to 10 seconds (daily activity). There was only a 1% difference in accuracy between the sampled and continuous tracking results and the MAE difference was below 1 bpm. If the HR were estimated once-a-minute, the reduction in the power consumption of the optical chain would be from 79.7% to 86.5% depending on the actual use-case. For longer sampling intervals the reduction could be even higher. However, it needs to be remembered that although the sampled HR accuracy is similar to the continuous HR algorithm, the availability of reliable HR estimation was lower than 80% during everyday activities. In conclusion, the experiments indicated that the proposed very-low-power semi-continuous HR algorithm could be used for tracking HR trends during 24/7 HR monitoring. The lower performance of the sampled HR during sport was caused mainly by the difficulty in correcting the motion errors. Nevertheless, the current results indicate that OHR monitoring could be developed for real long-term 24/7 usage without any significant loss of accuracy.

7 Discussion

7.1 Results versus objectives

The first objective was to develop an objective OHR evaluation methodology that allows OHR accuracy to be evaluated in various real life situations (Publications I-V)

This objective was achieved in the section “Evaluation framework” and Publications I-V. There has recently been a rapid increase in the number of available consumer OHR monitors, experimental OHR platforms and their use in different applications. However, there is a lack of common guidelines for OHR evaluation, which emphasizes the need for a unified evaluation methodology built on a scientific approach. The differences between the devices may be significant and therefore it is important to evaluate each device with a coherent evaluation methodology. The objective OHR evaluation methodology described in the “Evaluation framework” section was developed based on a comprehensive review of previous evaluation studies that identified the key factors affecting OHR with the addition of generic scientific methods.

The OHR evaluation process was divided into four basic stages: the design of the evaluation campaign including protocol design, selection of test subjects and proper reference devices; the execution of the evaluation campaign using important procedures for the operators who must strictly adhere to the instructions in the protocol; pre-processing the measured signals for evaluation paying particular attention to the time synchronization and the reference signal processing; and accuracy analyses highlighting the most suitable error estimation metric. The error metrics MAE and HR score < 10 bpm were selected because MAE is a generic error metric describing average disagreement, while HR score < 10 bpm is a practical metric showing the length of time when the estimated HR is close to the true reference. In addition, Bland-Altman plots were shown to be a useful graphical tool for comparison of the agreement between the two methods. The

proposed methodology for the objective evaluation of OHR accuracy was progressively applied in various consumer OHR evaluation studies in Publications I-V, which allowed a straightforward comparison of the key results of the evaluation studies.

The second objective was to evaluate the accuracy of selected high-end OHR devices during sports (Publications I, II and V)

This objective was achieved in Publications I, II and V. The HR accuracy of four consumer OHR devices, Mio Alpha, Mio Link, Scosche myRhythm and PulseOn were compared against ECG-based HR recorders during a controlled indoor-sport protocol and non-controlled outdoor exercises. A straightforward comparison of the results for each device and activity throughout the studies was achieved by applying a uniform evaluation methodology consisted of similar activities in the testing protocols, similar reference devices and identical statistical metrics. The representative number of test subjects for each study emphasized gender and age equality among the test group. The results demonstrated the high accuracy of the different OHR devices during endurance sport exercises, which ranged from 93.2 to 98.1%, corresponding to MAPE 1.9 – 6.8%. This is considerably more accurate than the similar sport exercise studies summarized in Table 4, which reported a much wider MAPE error range of 1.14 to 25.38%. There were only two studies (Stahl et al. 2016; Jo et al. 2016) showing similar MAPE error range from 3.28% to 9.8% during endurance exercises. The higher MAPE for the consumer OHR devices in the earlier studies may be caused by inappropriate operation of the devices, the design of the device itself, or the design and execution of the test protocol. The overall reliability achieved for all four of our tested devices in the sport protocols ranged from 86.6 to 95.4%, which agrees with the results of another study evaluating clinical OHR monitors in a combined (sport and simulated daily activity) protocol which had a reliability (HR score < 10 bpm) of 93.8% (Hendrikx et al. 2017). Slightly lower reliability in the range of 80.64 to 90.28 % was reported for OHR devices examined during a simulated everyday-activity protocol with short exercise extension (Pietilä et al. 2018). It must be noted that OHR accuracy may also vary between different OHR brands. This thesis used high-end devices but some cheaper devices have produced much worse results in other studies. The results of Publications I, II and V show very good accuracy for consumer OHR devices during endurance sport activities consisting mainly of rhythmic endurance exercises such as walking, running or cycling, provided that the devices are used properly according to the instructions and in the designed conditions.

The third objective was to evaluate beat-to-beat accuracy of a selected OHR device during sleep (Publication III)

This objective was achieved in Publication III. It was one of the first OHR evaluations focused on beat-to-beat accuracy. The PulseOn OHR was selected to compare the accuracy of OHR beat-to-beat detection with a standard ECG RRI recorder during sleep in normal non-controlled conditions at home. The number of missing, extra and correctly detected beats for the OHR device was analyzed with an automatic method which compared the reference beats within a specific time range with the detected beats. In this study, precise time synchronization of the signals was required to avoid major inaccuracies. The accuracy of the IBI estimation of corresponding beats was tested with time-aligned signals applying general statistical metrics. The results of all the analyses were calculated both for artifact-corrected and non-corrected data in order to show the impact and efficiency of the artifact-correction algorithm. In addition, estimation of HRV and sleep analysis parameters were compared for both the PPG and ECG beat detection methods. The results showed a high number of correctly detected beats during no motion activity. After applying artifact correction, the number of missing, extra and correctly detected beats for the PulseOn device were 0.43%, 0.72% and 99.57%, respectively. In a recent study, Pietilä et al. (Pietilä et al. 2018) evaluated the PulseOn device during a controlled daily-activity protocol that included activities with an increased amount of wrist motions. After the filtering artifacts had been applied, Pietila et al.'s results showed fewer correctly detected beats (76.2 – 90.3%) and higher relative numbers of extra (3.4 – 8.4%) and missing (5.9% – 15.4%) beats. Publication III also recognized a low mean bias between the OHR monitor and the ECG reference, 3.1 ms and (95% LoA) (−23.15, 23.83) ms for paired IBIs. In a former study performed during sleep, comparable results were reported that showed an even lower mean bias 0.05 ms, but a wider (95% LoA) (-35.7, 35.81) ms (Renevey et al. 2013). The evaluation study in Publication III demonstrated that OHR measurement can provide accurate beat-to-beat detection during sleep, which might be suitable as an input for HRV analyses. The results are in line with other studies which have focused on the accuracy of IBI estimations of OHR.

The fourth objective was to evaluate the accuracy of EE and VO_{2max} estimation based on OHR and mobile phone-based speed estimation (Publication V)

This objective was achieved in Publication V. The EE and VO_{2max} estimation accuracy based on OHR were among the first studies of this type, although several ECG-based HR VO_{2max} and EE evaluations are available nowadays. The accuracy of the estimation of EE and VO_{2max} based on OHR measurement was compared against the standard IC

reference method which analyses the gases in the subjects' breath. Maximal testing procedures adhering to the guidelines for performing sport exercise were used to estimate the reference EE, HR_{max} and VO_{2max} values. The EE estimation using OHR was measured during the reference measurement procedure in controlled laboratory conditions. VO_{2max} estimation based on OHR and mobile phone-based speed estimation was conducted according to a semi-controlled protocol of submaximal self-paced outdoor running. The outdoor running part was performed in regular winter training conditions, hence, it was a well-approximated testing protocol for real-use scenarios and provides a good benchmark for real outdoor training exercises. The EE estimation based on OHR was most accurate during heavy-intensity exercise with a MAPE of 6.7%. During light-intensity exercise the MAPE increased to 16.5%. The EE estimates based on OHR and IC had strong correlations during light-intensity exercise $\rho = 0.77$, while at a heavy-intensity exercise an even stronger correlation was observed, $r = 0.97$. These results are well in line with studies examining chest-strap-based HR for EE estimation which reported moderate ($r = 0.59$) to strong ($r = 0.975$) correlations during low and high intensity exercise, respectively (Erdogan et al. 2010; Robertson et al. 2015). However, previous studies evaluating EE estimation based on OHR reported higher MAPE rates at both intensity levels of exercise, 14.07% – 84.98% and 19.64% – 30.77%, for low-intensity and high-intensity exercise respectively (Dooley et al. 2017). A wide EE estimation error range was reported during ergo-cycle and resistance exercises, MAPE 21.13% — 75.15%, MAPE 42.69% - 57.02%, respectively (Boudreaux et al. 2018). In addition, recent studies have only shown weak to moderate correlation $r = 0.16 – 0.64$ between OHR-based EE estimation and IC (Wallen et al. 2016). The higher error rate and inaccuracy of the EE estimation in recent studies may be due to the inappropriate operation of the device. OHR- and speed-based VO_{2max} estimation during submaximal running can estimate VO_{2max} quite accurately. The MAPE was 5.2% for VO_{2max} when an individually measured HR_{max} parameter was used in the estimation. The MAPE increased slightly to 5.9% when the age-based HR_{max} value for VO_{2max} estimation was used. In addition, when the age-based HR parameter was used, strong to moderate correlations were observed for males and females, $r = 0.73$ and $r = 0.63$ respectively. These results are in line with the evaluations of VO_{2max} prediction using chest-strap HR, which reported correlations of $r = 0.872$ for males and $r = 0.477$ for females (Crouter et al. 2004). The OHR can be utilized to estimate the EE and VO_{2max} parameters in combination with the GPS speed and other basic user parameters as long as appropriate algorithms incorporating physiological modelling are applied.

The fifth objective was to design and evaluate a low-power approach to OHR estimation for everyday use (Publication IV).

This objective was achieved in Publication IV. The semi-continuous HR "Sampled HR" estimation algorithm was designed for a low-power approach to OHR estimation during long-term daily monitoring. The algorithm's performance using MAE, reliability and accuracy metrics was evaluated against the average real HR measurements on various datasets including indoor and outdoor exercise, everyday activity and sleep. In addition, two special metrics were defined to measure the semi-continuous HR performance: estimation delay, and reliable estimation delay. The error statistics were compared with continuous HR estimations. The algorithm implementation and evaluation was executed offline only in a computer environment. The expected relative reduction of power consumption against continuous HR monitoring was also calculated based on the results obtained from off-line simulations. An average estimation delay of reliable sampled HR variation was from 7.1 seconds in sport to 10 seconds during everyday activities. The MAE and MAPE varied throughout all the datasets in the ranges of 1.3 to 4.5 bpm and 2.4 to 4.6%, respectively. A similar study evaluating minute-by-minute HR estimations in free living conditions had a slightly higher MAPE of 9.17% (J. Lee et al. 2016), although the algorithm for HR estimation was not presented The simulations in Publication IV showed a possible reduction in the power consumption of the optical chain from 79.7% to 86.5%. However, it should be noted that with the sampled HR algorithm, HR estimation was only available between 94.4% and 77.1% of execution time, which is the price to be paid for not running the HR calculation continuously. It was concluded that semi-continuous HR estimation can provide high estimation accuracy and a potential increase in the device's lifetime due to the reduction in power consumption.

7.2 Impacts of the studies in their research fields

Until now there have not been any unified guidelines or standards available which can be generalized for the OHR evaluation process. The 'evaluation framework' in this research defines the key topics of any systematic and objective OHR evaluation and is essential for making valid comparisons between studies. The definition of the evaluation methodology reflects current needs and takes into account the main factors affecting the PPG signal quality. The defined processes and methods are applicable both for consumer OHR monitors and experimental research platforms. Moreover, the procedures in the framework can be also used for the comparison and evaluation of various other HR

estimations. Bassett et al (Bassett et al. 2012) have already emphasized the importance of being able to directly compare the accuracy of evaluations of wearable sensors' signals (e.g. for HR) which are used to derive estimates for further parameters (e.g. EE, VO_{2max}). Some of the factors that can affect the PPG signal may be put down to inaccuracies in the HR estimation. Thus, all researchers should perform 'unit calibration' and check the accuracy of signals and variables provided by the sensors before executing the chosen application (Bassett et al. 2012). The methodology described here has been developed and incrementally applied in the presented evaluations.

The studies in Publications I, II and V showed that consumer wrist-worn OHR devices can provide reliable results for HR estimation during indoor rhythmic sports as long as the user follows the manufacturer's instructions. A number of other studies evaluating wrist-worn OHR in sport conditions that have been published since this thesis was begun report the similar results for reliability and accuracy (Spierer et al. 2015; Jo et al. 2016; Stahl et al. 2016; Hendrikx et al. 2017). The OHR accuracy during maximal HR testing and outdoor sports conditions have not yet been widely studied. Publications II and V showed good OHR reliability during both outdoor rhythmic sports and maximal HR testing. Publication I showed significantly better HR accuracy during non-rhythmic movement with the sensor placed on the forearm rather than the wrist, and this is confirmed by a study examining PPG signal quality on different body locations (Maeda et al. 2011).

Few studies have focused on the long-term IBI accuracy of wrist-worn OHR devices. Schafer and Vager (Schäfer & Vagedes 2013) in their review of HRV demonstrated agreement between finger-based pulse rate and HRV. In Pub III, the high proportion of correctly-detected beats and the small difference between the OHR and ECG IBI values supports the assumption that HRV derived from OHR can be utilized during sleep or other motionless conditions. The results obtained in Pub III, especially those for mean bias and the corresponding limits of agreement in the BA plots, are supported by a similar study evaluating OHR during sleep (Renevey et al. 2013). All of the key findings in Pub III support the use of OHR-based IBI in any further clinical, sleep, behavioral or other studies which need unobtrusive long-term HR monitoring during low motion.

Pub V presented OHR with a combination of appropriate physiological modelling algorithms, especially during rhythmic sports. These allowed the EE and VO_{2max} to be estimated accurately. The correlations between the OHR-based EE and the VO_{2max} estimations agree with the chest-strap HR-based predictions for the same parameters in previous studies (Crouter et al. 2004; Erdogan et al. 2010; Robertson et al. 2015). Moreover, Pub V also proved that there is a slight error in the VO_{2max} estimations based on OHR

and GPS speed during submaximal running in challenging outdoor conditions. Although the chest-strap HR-based VO_{2max} estimations have shown a lower error using individual true HR_{max} measurements (Crouter et al. 2004; Montgomery et al. 2009; Erdogan et al. 2010), no significant difference was found between using true measured HR_{max} or age-based HR_{max} for OHR-based VO_{2max} calculations. Nonetheless, the results in Publication V demonstrate notably higher accuracy for OHR-based EE prediction than similar evaluations performed previously (Wallen et al. 2016; Dooley et al. 2017; Boudreaux et al. 2018). Similar error rates in EE estimation were only presented with clinical OHR monitor evaluation (Hendrikx et al. 2017). The variation in EE estimation accuracy might be caused by differences in the calculation methodology of the different brands.

A novel approach to reducing power consumption in long-term HR monitoring was presented and verified on various datasets in Publication IV. The semi-continuous algorithm was highly accurate and reliable, and had a low error rate in HR detection making it a viable alternative to continuous monitoring. The new parameters, estimation delay and reliability, show how effective the semi-continuous algorithm is. They were defined and utilized during the simulations of power reduction efficiency and the results show possible savings in power consumption due to switching off the optical chain. Therefore, Publication IV can be seen as a simple framework for evaluating semi-continuous HR algorithms, one which didn't exist before.

7.3 Limitations of the studies

In Publications I, II, III, IV and V the demographic parameters of the test subjects were too homogenous to evaluate the impact of different skin color and possible skin tissue structure changes in old age. The volunteer test subjects in the listed publications were healthy, young (20-45 years old) Caucasians with Fitzpatrick skin types I, II and III. This cohort of test subjects had very good skin perfusion and provided a high-quality input signal for the OHR sensors. It is therefore possible that the final HR estimation error may be lower than in other population samples.

In Publications I, II, III, IV and V the order of activities in the protocol was not randomized. This may affect the accuracy of the HR results because the latter activities in the protocol may benefit from improved perfusion caused by any elevated physical activity performed earlier in the protocol.

In Publications I and II, the controlled protocol conditions were limited to indoor laboratory protocols, but there were also non-controlled outdoor protocols. However, when the outdoor protocol was performed, the effects of ambient light (sunshine) or low skin perfusion (cold weather) could not be tested and controlled. In addition, the test protocols were predominantly rhythmic sport activities (e.g. running, walking, or cycling). There were no non-rhythmic sports (e.g. ball games, gym exercise) included in the protocols.

In Publication III, which evaluated the accuracy of IBI, the sample was small (only 10 subjects). In Publication III, the used reference device recorded only RR intervals and therefore it was not possible to verify if all of R-peaks were detected correctly. Therefore, the quality of the reference data in this study could not be verified. However, the accuracy of RR interval detection was verified and provided on manufactures pages and hence the quality was likely high.

In Publications III and V only one device was tested. This was due to practical reasons, as getting the test subjects to wear multiple devices in the laboratory, during outdoor exercise and during sleep would have been difficult to oversee and control. This may limit the generalizability of the obtained results.

In Publication VI, a semi-continuous HR algorithm was evaluated in offline simulations. The proposed algorithm was not actually used with a real device so there is no reference measurement for actual savings in power consumption, nor a comparison of the accuracy of real-time device HR estimations with continuous HR estimations. In addition, the offline simulation for the semi-continuous HR approach only used one set of optimal sampling-interval parameters.

In Publications II, III and V, BA Plots were used to display the HR, IBI, EE and VO_{2max} error without prior inspection of the type of bias (fixed or proportional). However, the BA Plot was only used as a visual representation of the error, and not as a method to estimate the limits of agreement; an approach which has often been used in other studies. In addition, both methods were measured in the same units and the results were not used to calibrate the devices. Hence, the BA plot may be treated as a valid method for giving a visual representation of the error regardless of the type of bias.

7.4 Directions for future research

The main focus of the studies in Publications I and II was to evaluate the accuracy of the HR estimation of HR sensors during indoor rhythmic endurance sports. The number of recorded measurements in outdoor conditions was limited in Publication II. Furthermore, little research and development has been done to give reliable results during non-rhythmic motion. Thus, it is important to continue performing evaluations of the accuracy of OHR sensors during non-rhythmic sport activities. In addition, more controlled outdoor-testing protocols are required to investigate the influence of ambient light from direct sunshine on a PPG signal, or (as is common in Scandinavia) the poor perfusion caused by a cold environment.

The lower HR estimation accuracy for darker skin colors has been demonstrated in evaluation studies done by Spierer et al (Spierer et al. 2015). Thus, there should be more evaluations with darker-skinned test subjects under various test conditions. Moreover, it would be beneficial to combine testing the effect of skin color on OHR accuracy with the impact of using different wavelength light sources.

The long-term monitoring of out-patients typically requires unobtrusive, user-friendly and reliable devices with zero maintenance (Korhonen et al. 2003) and wrist-worn OHR trackers are perfect for this. Publication III has already shown good IBI detection accuracy with OHR, making it suitable for HRV monitoring during sleep in non-controlled conditions with young healthy adults. Relatively high IBI and HR estimation accuracy with OHR has been also shown during a simulated everyday-life protocol (Pietilä et al. 2018). In the future it will be really important to perform long-term OHR evaluations in 24/7 non-controlled protocols to demonstrate their usefulness in the long-term monitoring of remote patients. Obviously, the elderly are the most likely target group for this technology in the future. PPG waveform shape and amplitude changes are related to increasing age (Allen & Murray 2003) and arterial stiffness (Brillante et al. 2008) and these changes can cause various inaccuracies in the measurements. Thus, future evaluations of OHR-based HR and IBI estimation accuracy should be conducted with older users. Studies should also be made with cardiovascular disease patients, or with the sufferers of other diseases such as diabetes and neurological disorders. Such studies would improve the reliability of this technology. Any future evaluations of OHR sensors, whether for research, sport and fitness, or medicine, should be systematic and comparable. If future researchers use the “Evaluation framework” presented here, it will be easier to perform objective and comparable evaluations of any new sensors and algorithms.

8 Conclusions

This thesis has presented an objective framework for the evaluation of optical wearable HR sensors focused on practical target applications. An evaluation framework describing the key procedures for an OHR evaluation methodology has been developed to standardize and to unify the methods used in future OHR evaluations. The framework was developed by identifying the main factors affecting OHR measurements, based on a comprehensive literature review and on the practical aspects of OHR technology. The methodology was directly applied in five OHR-sensor evaluation studies. The primary aim of these studies was to explore the accuracy of HR and IBI detection. The accuracy of EE and VO_{2max} parameters based on OHR was also verified. An approach to semi-continuous HR estimation enabling lower power consumption was designed and evaluated against continuous HR measurements. The following key findings can be drawn from the studies included in this thesis:

- A unified approach based on scientific methods is required for the evaluation of OHR sensors so that the results of different studies can easily be compared.
- It has been proved that OHR sensors can provide reliable and accurate HR estimates during rhythmic sport activities when compared to ECG-based standard devices. This has been shown in both controlled indoor and non-controlled outdoor conditions.
- During sleep (e.g. when the test subject exhibits minimal movement) OHR can reliably detect beats from PPG and accurately estimate IBI.
- IBI based on OHR can be utilized to calculate HRV and sleep quality parameters.
- EE and VO_{2max} may be accurately estimated on the basis of the OHR measurement in combination with appropriate physiological modelling and the GPS-measured speed during rhythmic sports.

- Semi-continuous HR estimation based on OHR technology provides a way to significantly reduce the power consumption. The method has acceptable accuracy when compared with continuous HR detection methods during sport, sleep or daily activities. However, the savings in power consumption are made at the expense of HR availability.

References

Achten, J. & Jeukendrup, A.E. (2003). Heart rate monitoring: applications and limitations, Sports medicine (Auckland, N.Z.), Vol. 33(7), pp. 517-538.

Agewall, S., Tjessem, L.H., Rossignol, P., Zannad, F., Atar, D., Lamiral, Z., Machu, J.L., Dickstein, K., Kjekshus, J., von Lueder, T.G., Girerd, N. & High Risk Myocardial Infarction Database Initiative investigators (2017). Heart rate prediction of outcome in heart failure following myocardial infarction depend on heart rhythm status an analysis from the high-risk myocardial infarction database initiative, International journal of cardiology, Vol. 249, pp. 274-281.

Allen, J. (2007). Photoplethysmography and its application in clinical physiological measurement, Physiological Measurement, Vol. 28(3), pp. R1-R39. Available (accessed 10.5.2018): http://iopscience.iop.org/article/10.1088/0967-3334/28/3/R01/meta.

Allen, J. & Murray, A. (2003). Age-related changes in the characteristics of the photoplethysmographic pulse shape at various body sites, Physiological Measurement, Vol. 24(2), pp. 297-307.

Altini, M. (2015). Personalization of energy expenditure and cardiorespiratory fitness estimation using wearable sensors in supervised and unsupervised free-living conditions, Technische Universiteit Eindhoven, 224 p. Available: https://pure.tue.nl/ws/files/11369233/20151215_Altini.pdf.

Altman, D.G. & Bland, J.M. (1983). Measurement in Medicine: The Analysis of Method Comparison Studies, Journal of the Royal Statistical Society.Series D (The Statistician), Vol. 32(3), pp. 307-317.

Atkinson, G. & Nevill, A.M. (1998). Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine, Sports medicine (Auckland, N.Z.), Vol. 26(4), pp. 217-238.

Aubert, A.E., Seps, B. & Beckers, F. (2003). Heart rate variability in athletes, Sports medicine (Auckland, N.Z.), Vol. 33(12), pp. 889-919.

Bannach, D., Amft, O. & Lukowicz, P. (2009). Automatic Event-Based Synchronization of Multimodal Data Streams from Wearable and Ambient Sensors, 2009 European Conference on Smart Sensing and Context, Guildford, 16 - 28 Sep., Springer, Berlin, Heidelberg, pp. 135-148.

Bassett, D.R.,Jr & Howley, E.T. (2000). Limiting factors for maximum oxygen uptake and determinants of endurance performance, Medicine and science in sports and exercise, Vol. 32(1), pp. 70-84.

Bassett, D.R.,Jr, Rowlands, A. & Trost, S.G. (2012). Calibration and validation of wearable monitors, Medicine and science in sports and exercise, Vol. 44(1 Suppl 1), pp. S32-S38.

Bay, N.S. & Bay, B. (2010). Greek anatomist herophilus: the father of anatomy, Anatomy & Cell Biology, Vol. 43(4), pp. 280-283.

Bedford, D.E. (1951). The ancient art of feeling the pulse, British heart journal, Vol. 13(4), pp. 423-437.

Benedetto, S., Caldato, C., Bazzan, E., Greenwood, D.C., Pensabene, V. & Actis, P. (2018). Assessment of the Fitbit Charge 2 for monitoring heart rate, PLoS ONE, Vol. 13(2). Available (accessed 10.5.2018): https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0192691/.

Bland, J.M. & Altman, D.G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement, Lancet (London, England), Vol. 1(8476), pp. 307-310.

Boudreaux, B.D., Hebert, E.P., Hollander, D.B., Williams, B.M., Cormier, C.L., Naquin, M.R., Gillan, W.W., Gusew, E.E. & Kraemer, R.R. (2018). Validity of Wearable Activity Monitors during Cycling and Resistance Exercise, Medicine and science in sports and exercise, Vol. 50(3), pp. 624-633.

Brillante, D.G., O'Sullivan, A.J. & Howes, L.G. (2008). Arterial stiffness indices in healthy volunteers using non-invasive digital photoplethysmography, Blood pressure, Vol. 17(2), pp. 116-123.

Bronzino, J.D. (1995). The biomedical engineering handbook, CRC Press, Boca Raton, Florida, 2862 p.

Bureau Veritas, Wearable Technology Solutions, web page. Available (accessed 9.5.2018): http://www.bureauveritas.com/services+sheet/wearable+technology.

Cambridge University Press, Cambridge University Press Meaning of “evaluation” in the English Dictionary, web page. Available (accessed 9.5.2018): https://dictionary.cambridge.org/dictionary/english/evaluation.

Challoner, A.V. & Ramsay, C.A. (1974). A photoelectric plethysmograph for the measurement of cutaneous blood flow, Physics in Medicine and Biology, Vol. 19(3), pp. 317-328.

Charlot, K., Cornolo, J., Borne, R., Brugniaux, J.V., Richalet, J.P., Chapelot, D. & Pichon, A. (2014). Improvement of energy expenditure prediction from heart rate during running, Physiological Measurement, Vol. 35(2), pp. 253-266.

Claes, J., Buys, R., Avila, A., Finlay, D., Kennedy, A., Guldenring, D., Budts, W. & Cornelissen, V. (2017). Validity of heart rate measurements by the Garmin Forerunner 225 at different walking intensities, Journal of medical engineering & technology, Vol. 41(6), pp. 480-485.

Clinical trial information leaflet and consent (2016). Valvira, Helsinki, 7 p.

Conrad, M.C. (1971). Functional Anatomy of the Circulation to the Lower Extremities: With Color Atlas, Year Book Medical, Chicago, 190 p.

Crouter, S.E., Albright, C. & Bassett, D.R.,Jr (2004). Accuracy of polar S410 heart rate monitor to estimate energy cost of exercise, Medicine and science in sports and exercise, Vol. 36(8), pp. 1433-1439.

Cui, W., Ostrander, L.E. & Lee, B.Y. (1990). In vivo reflectance of blood and tissue as a function of light wavelength, IEEE Transactions on Biomedical Engineering, Vol. 37(6), pp. 632-639.

Dassel, A.C.M., Graaff, R., Sikkema, M., Meijer, A., Zijlstra, W.G. & Aarnoudse, J.G. (1995). Reflectance pulse oximetry at the forehead improves by pressure on the probe, Journal of clinical monitoring, Vol. 11(4), pp. 237-244.

Delgado-Gonzalo, R., Renevey, P., Lemkaddem, A., Lemay, M., Sola, J., Korhonen, I. & Bertschi, M. (2018). Physical Activity, in: Tamura, T. & ChenWenxi (ed.), Seamless Healthcare Monitoring, 1st ed., Springer International Publishing, Cham, pp. 413-455.

de Zambotti, M., Baker, F.C., Willoughby, A.R., Godino, J.G., Wing, D., Patrick, K. & Colrain, I.M. (2016). Measures of Sleep and Cardiac Functioning During Sleep Using a

Multi-Sensory Commercially–Available Wristband in Adolescents: Wearable Technology to Measure Sleep and Cardiac Functioning, Physiology & Behavior, Vol. 158, pp. 143-149.

Dooley, E.E., Golaszewski, N.M. & Bartholomew, J.B. (2017). Estimating Accuracy at Exercise Intensities: A Comparative Study of Self-Monitoring Heart Rate and Physical Activity Wearable Devices, JMIR mHealth and uHealth, Vol. 5(3). Available (accessed 10.5.2018): https://mhealth.jmir.org/2017/3/e34/.

Dresher, R.P. & Mendelson, Y. (2006). Reflectance Forehead Pulse Oximetry: Effects of Contact Pressure During Walking, 2006 International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), New York, 30 Aug. - 3 Sept., IEEE, pp. 3529-3532.

Einthoven, W. (1895). Ueber die Form des menschlichen Electrocardiogramms, Archiv für die gesamte Physiologie des Menschen und der Tiere, Vol. 60(3), pp. 101-123.

Erdogan, A., Cetin, C., Karatosun, H. & Baydar, M.L. (2010). Accuracy of the Polar S810i(TM) Heart Rate Monitor and the Sensewear Pro Armband(TM) to Estimate Energy Expenditure of Indoor Rowing Exercise in Overweight and Obese Individuals, Journal of sports science & medicine, Vol. 9(3), pp. 508-516.

Etikan, I., Musa, S.A. & Alkassim, R.S. (2016). Comparison of Convenience Sampling and Purposive Sampling, American Journal of Theoretical and Applied Statistics, Vol. 5(1), pp. 1-4. Available (accessed 10.5.2018): http://www.sciencepublishinggroup.com/journal/paperinfo?journalid=146&doi=10.11648/j.ajtas.20160501.11.

Ethical principles of research in the humanities and social and behavioural sciences and proposals for ethical review (2009). TENK, Helsinki, 17 p.

Fallet, S. & Vesin, J.M. (2017). Robust heart rate estimation using wrist-type photoplethysmographic signals during physical exercise: an approach based on adaptive filtering, Physiological Measurement, Vol. 38(2), pp. 155-170.

Fallow, B.A., Tarumi, T. & Tanaka, H. (2013). Influence of skin type and wavelength on light wave reflectance, Journal of clinical monitoring and computing, Vol. 27(3), pp. 313-317.

Firstbeat Technologies Ltd., Bodyguard 2 and Electrodes Ambu BlueSensor L, web page. Available (accessed 10.5.2018): http://shop.firstbeat.fi/all-products/.

Firstbeat Technologies Ltd., Stress and Recovery Analysis Method Based on 24-hour Heart Rate Variability, web page. Available (accessed 10.5.2018): https://assets.firstbeat.com/firstbeat/uploads/2015/11/Stress-and-recovery_white-paper_20145.pdf.

Fitzpatrick, T.B. (1988). The validity and practicality of sun-reactive skin types I through VI, Archives of Dermatology, Vol. 124(6), pp. 869-871.

Floyer, S.J. (1707). The Physician's Pulse-Watch; or, an Essay to Explain the Old Art of Feeling the Pulse, and to Improve it by Help of the Pulse Watch, S. Smith and B. Walford, London, 13 p.

Floyer, S.J. (1710). The Pulse Watch, J. Nicholson, W. Taylor and J. H. Clements, London.

Fohr, T., Tolvanen, A., Myllymaki, T., Jarvela-Reijonen, E., Peuhkuri, K., Rantala, S., Kolehmainen, M., Korpela, R., Lappalainen, R., Ermes, M., Puttonen, S., Rusko, H. & Kujala, U.M. (2017). Physical activity, heart rate variability-based stress and recovery, and subjective stress during a 9-month study period, Scandinavian Journal of Medicine & Science in Sports, Vol. 27(6), pp. 612-621.

Franklin, B.A. & Balady, G.J. (2000). ACSM's guidelines for exercise testing and prescription, 6th ed. Lippincott Williams & Wilkins, Philadelphia (Pa.), 368 p.

Fraser, C.G. (2001). Comprehensive summary of within-subject biological variation that relates to other sources of variation in laboratory testing, in: Fraser, C.G. (ed.), Biological Variation: From Principles to Practice, AACC Press, Washington DC, 150 p.

Garmin Ltd., Forerunner® 610 and Soft Strap Premium Heart Rate Monitor, web page. Available (accessed 10.5.2018): https://buy.garmin.com/en-US/.

Gorny, W.A., Liew, J.S., Tan, S.C. & Mueller-Riemenschneider, F. (2017). Fitbit Charge HR Wireless Heart Rate Monitor: Validation Study Conducted Under Free-Living Conditions, JMIR Mhealth Uhealth, Vol. 5(10). Available (accessed 10.5.2018): https://mhealth.jmir.org/2017/10/e157/.

Grabovskis, A., Marcinkevics, Z., Rubins, U. & Kviesis-Kipge, E. (2013). Effect of probe contact pressure on the photoplethysmographic assessment of conduit artery stiffness, Journal of Biomedical Optics, Vol. 18(2). Available (accessed 10.5.2018): https://doi.org/10.1117/1.JBO.18.2.027004.

Hallman, D.M., Mathiassen, S.E. & Lyskov, E. (2015). Long-Term Monitoring of Physical Behavior Reveals Different Cardiac Responses to Physical Activity among Subjects with and without Chronic Neck Pain, BioMed Research International, Vol. 2015. Available (accessed 10.5.2018): https://www.hindawi.com/journals/bmri/2015/907482/.

Haykin, S. (2001). Adaptive Filter Theory, 4th ed. Prentice Hall, New Jersey, 920 p.

Hendrikx, J., Ruijs, S.L., Cox, G.L., Lemmens, M.P., Schuijers, G.E. & Goris, H.A. (2017). Clinical Evaluation of the Measurement Performance of the Philips Health Watch: A Within-Person Comparative Study, JMIR Mhealth Uhealth, Vol. 5(2). Available (accessed 10.5.2018): https://mhealth.jmir.org/2017/2/e10/.

Hertzman, A.B. (1938). The blood supply of various skin areas as estimated by the photoelectric plethysmograph, American Journal of Physiology-Legacy Content, Vol. 124(2), pp. 328-340.

Hertzman, A.B. (1937). Photoelectric Plethysmography of the Fingers and Toes in Man, Proceedings of the Society for Experimental Biology and Medicine, Vol. 37(3), pp. 529-534.

Holter, N.J. (1961). New Method for Heart Studies, Science, Vol. 134(3486), pp. 1214-1220.

Hopkins, W.G. (2004). Bias in Bland-Altman but not Regression Validity Analyses, Sportscience, Vol. 8, pp. 42-46. Available (accessed 9.5.2018): http://sportsci.org/jour/04/wghbias.htm.

Hopkins, W.G., Marshall, S.W., Batterham, A.M. & Hanin, J. (2009). Progressive statistics for studies in sports medicine and exercise science, Medicine and science in sports and exercise, Vol. 41(1), pp. 3-13.

Hwang, S., Seo, J.O., Jebelli, H. & Lee, S.H. (2016). Feasibility analysis of heart rate monitoring of construction workers using a photoplethysmography (PPG) sensor embedded in a wristband-type activity tracker, Automation in Construction, Vol. 71, pp. 372-381.

ISO 14155 (2011). Clinical investigation of medical devices for human subjects — Good clinical practice, International Organization for Standardization, Geneva, 58 p.

ISO 80601-2-61 (2011). Medical electrical equipment — Part 2-61: Particular requirements for basic safety and essential performance of pulse oximeter equipment, International Organization for Standardization, Geneva, 84 p.

Jayasree, V., Sandhya, T. & Radhakrishnan, P. (2008). Non-invasive Studies on Age Related Parameters Using a Blood Volume Pulse Sensor, Measurement Science Review, Vol. 8, pp. 82-86.

Jeukendrup, A.E. & Wallis, G.A. (2005). Measurement of substrate oxidation during exercise by means of gas exchange measurements, International Journal of Sports Medicine, Vol. 26 Suppl 1, pp. S28-S37.

Jo, E., Lewis, K., Directo, D., Kim, M.J. & Dolezal, B.A. (2016). Validation of Biofeedback Wearables for Photoplethysmographic Heart Rate Tracking, Journal of sports science & medicine, Vol. 15(3), pp. 540-547.

Joint Committee for Guides in Metrology (2008). International vocabulary of metrology – Basic and general concepts and associated terms (VIM), 3rd ed. Bureau International des Poids et Mesures, 91 p.

Jones, D.P. (1987). Medical electro-optics: measurements in the human microcirculation, Physics in Technology, Vol. 18(2), pp. 79-85.

Kaikkonen, P., Lindholm, H. & Lusa, S. (2017). Physiological Load and Psychological Stress During a 24-hour Work Shift Among Finnish Firefighters, Journal of occupational and environmental medicine, Vol. 59(1), pp. 41-46.

Kamal, A.A., Harness, J.B., Irving, G. & Mearns, A.J. (1989). Skin photoplethysmography — a review, Computer methods and programs in biomedicine, Vol. 28(4), pp. 257-269.

Keytel, L.R., Goedecke, J.H., Noakes, T.D., Hiiloskorpi, H., Laukkanen, R., van der Merwe, L. & Lambert, E.V. (2005). Prediction of energy expenditure from heart rate monitoring during submaximal exercise, Journal of sports sciences, Vol. 23(3), pp. 289-297.

Khushhal, A., Nichols, S., Evans, W., Gleadall-Siddall, D., Page, R., O'Doherty, A., Carroll, S., Ingle, L. & Abt, G. (2017). Validity and reliability of the Apple Watch for measuring heart rate during exercise, Sports Medicine International Open, Vol. 1(06), pp. E206-E211. Available (accessed 10.5.2018): https://www.thieme-connect.com/products/ejournals/abstract/10.1055/s-0043-120195.

Kim, J., Lee, T., Kim, J. & Ko, H. (2015). Ambient light cancellation in photoplethysmogram application using alternating sampling and charge redistribution technique, 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Milan, 25 - 29 Aug., IEEE, pp. 6441-6444.

Kingsley, M., Lewis, M.J. & Marson, R.E. (2005). Comparison of Polar 810s and an ambulatory ECG system for RR interval measurement during progressive exercise, International Journal of Sports Medicine, Vol. 26(1), pp. 39-44.

Korhonen, I. (1997). Methods for the analysis of short-term variability of heart rate and blood pressure in frequency domain, VTT, 96 p. Available: http://www.vtt.fi/inf/pdf/publications/1997/P316.pdf.

Korhonen, I., Parkka, J. & Gils, M.V. (2003). Health monitoring in the home of the future, IEEE Engineering in Medicine and Biology Magazine, Vol. 22(3), pp. 66-73.

Kristiansen, J., Korshoj, M., Skotte, J.H., Jespersen, T., Sogaard, K., Mortensen, O.S. & Holtermann, A. (2011). Comparison of two systems for long-term heart rate variability monitoring in free-living conditions - a pilot study, Biomedical engineering online, Vol. 10. Available (accessed 10.5.2018): https://biomedical-engineering-online.biomedcentral.com/articles/10.1186/1475-925X-10-27.

Kroll, R.R., Boyd, J.G. & Maslove, D.M. (2016). Accuracy of a Wrist-Worn Wearable Device for Monitoring Heart Rates in Hospital Inpatients: A Prospective Observational Study, Journal of medical Internet research, Vol. 18(9). Available (accessed 10.5.2018): https://www.jmir.org/2016/9/e253/.

Kumazawa, T., Kobayashi, M. & Takagi, K. (1964). A Plethysmographic Study of the Human Skin Under various Environmental Conditions, The Japanese journal of physiology, Vol. 14, pp. 354-364.

Lee, B., Kee, Y., Han, J. & Yi, W.J. (2011). Adaptive comb filtering for motion artifact reduction from PPG with a structure of adaptive lattice IIR notch filter, 2011 33rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Boston, 30. Aug. - 3 Sept., IEEE, pp. 7937-7940.

Lee, C.M. & Gorelick, M. (2011). Validity of the Smarthealth Watch to Measure Heart Rate During Rest and Exercise, Measurement in Physical Education and Exercise Science, Vol. 15(1), pp. 18-25.

Lee, J., An, H., Kang, S. & Kim, Y. (2016). Examining the Validity of Fitbit Charge HR for Measuring Heart Rate in Free-Living Conditions, University of Nebraka at Omaha, Available: https://digitalcommons.unomaha.edu/pahppresentations/15.

Lemay, M., Bertschi, M., Sola, J., Renevey, P., Parak, J. & Korhonen, I. (2014). Application of Optical Heart Rate Monitoring, in: Sazonov, E. & Neuman, M.R. (ed.), Werable Sensors: Fundamentals, Implementation and Applications, Elsevier; Imprint: Academic Press, Oxford, pp. 105-129.

Levine, J.A. (2005). Measurement of energy expenditure, Public health nutrition, Vol. 8(7A), pp. 1123-1132.

Lindberg, L.G. & Oberg, P.A. (1991). Photoplethysmography. Part 2. Influence of light source wavelength, Medical & biological engineering & computing, Vol. 29(1), pp. 48-54.

Linnet, K. (1993). Evaluation of regression procedures for methods comparison studies, Clinical chemistry, Vol. 39(3), pp. 424-432.

Ludbrook, J. (2010). Confidence in Altman-Bland plots: a critical review of the method of differences, Clinical and experimental pharmacology & physiology, Vol. 37(2), pp. 143-149.

Ludbrook, J. (2010). Linear regression analysis for comparing two measurers or methods of measurement: but which regression? Clinical and experimental pharmacology & physiology, Vol. 37(7), pp. 692-699.

Maeda, Y., Sekine, M. & Tamura, T. (2011). Relationship Between Measurement Site and Motion Artifacts in Wearable Reflected Photoplethysmography, Journal of medical systems, Vol. 35(5), pp. 969-976.

Maeda, Y., Sekine, M., Tamura, T. & Mizutani, K. (2013). The Effect of Contact Pressure to the Photoplethysmographic Sensor During Walking, Transactions of Japanese Society for Medical and Biological Engineering, Vol. 51, pp. 307-307.

Maeda, Y., Sekine, M. & Tamura, T. (2011). The advantages of wearable green reflected photoplethysmography, Journal of medical systems, Vol. 35(5), pp. 829-834.

Maeda, Y., Sekine, M., Tamura, T., Moriya, A., Suzuki, T. & Kameyama, K. (2008). Comparison of reflected green light and infrared photoplethysmography, 2008 30th

Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Vancouver, 20 - 25 Aug. , IEEE, pp. 2270-2272.

Malik, M. (1996). Heart Rate Variability, Annals of Noninvasive Electrocardiology, Vol. 1(2), pp. 151-181.

Marieb, E.N. (2006). Essentials of human anatomy & physiology, 8th ed. Pearson/Benjamin Cummings, San Francisco, 606 p.

Martinez-Mesa, J., Gonzalez-Chica, D.A., Bastos, J.L., Bonamigo, R.R. & Duquia, R.P. (2014). Sample size: how many participants do I need in my research? Anais Brasileiros de Dermatologia, Vol. 89(4), pp. 609-615.

Mateo, J. & Laguna, P. (2003). Analysis of heart rate variability in the presence of ectopic beats using the heart timing signal, IEEE Transactions on Biomedical Engineering, Vol. 50(3), pp. 334-343.

Mendelson, Y. & Ochs, B.D. (1988). Noninvasive pulse oximetry utilizing skin reflectance photoplethysmography, IEEE Transactions on Biomedical Engineering, Vol. 35(10), pp. 798-805.

Merilahti, J., Parkka, J., Antila, K., Paavilainen, P., Mattila, E., Malm, E.J., Saarinen, A. & Korhonen, I. (2009). Compliance and technical feasibility of long-term health monitoring with wearable and ambient technologies, Journal of telemedicine and telecare, Vol. 15(6), pp. 302-309.

Monach, P.A. (2012). Repeating tests: different roles in research studies and clinical medicine, Biomarkers in medicine, Vol. 6(5), pp. 691-703.

Montgomery, P.G., Green, D.J., Etxebarria, N., Pyne, D.B., Saunders, P.U. & Minahan, C.L. (2009). Validation of heart rate monitor-based predictions of oxygen uptake and energy expenditure, Journal of strength and conditioning research, Vol. 23(5), pp. 1489-1495.

Mutikainen, S., Helander, E., Pietilä, J., Korhonen, I. & Kujala, U.M. (2014). Objectively measured physical activity in Finnish employees: a cross-sectional study, BMJ Open, Vol. 4(12). Available (accessed 10.5.2018): https://bmjopen.bmj.com/content/4/12/e005927.

Myllymaki, T., Rusko, H., Syvaoja, H., Juuti, T., Kinnunen, M.L. & Kyrolainen, H. (2012). Effects of exercise intensity and duration on nocturnal heart rate variability and sleep quality, European journal of applied physiology, Vol. 112(3), pp. 801-809.

Nieman, D. (2011). Exercise Testing & Prescription, 7th ed. McGraw-Hill, New York, 652 p.

Nijboer, J.A., Dorlas, J.C. & Mahieu, H.F. (1981). Photoelectric plethysmography — some fundamental aspects of the reflection and transmission method, Clinical physics and physiological measurement : an official journal of the Hospital Physicists' Association, Deutsche Gesellschaft fur Medizinische Physik and the European Federation of Organisations for Medical Physics, Vol. 2(3), pp. 205-215.

Nilsson, L., Goscinski, T., Johansson, A., Lindberg, L.G. & Kalman, S. (2006). Age and gender do not influence the ability to detect respiration by photoplethysmography, Journal of clinical monitoring and computing, Vol. 20(6), pp. 431-436.

Nummela, A. (2007). Aerobisen kestävyyden suorat mittausmenetelmät [Direct aerobic endurance measurement methods], in: Keskinen, K., Häkkinen, K. & Kallinen, M. (ed.), Kuntotestauksen käsikirja [Fitness testing handbook], 2nd ed., Liikuntatieteellinen Seura [Finnish Society of Sport Sciences], Helsinki, pp. 64-78.

O'Connor, D.P., Mahar, M.T., Laughlin, M.S. & Jackson, A.S. (2011). The Bland-Altman method should not be used in regression cross-validation studies, Research quarterly for exercise and sport, Vol. 82(4), pp. 610-616.

Olenick, A.A., Haile, L. & Dixon, C.B. (2015). Validation of the Mio Alpha Heart Rate Monitor during Graded Exercise Testing in Trail Runners, International Journal of Exercise Science: Conference Proceedings, Vol. 9(3). Available (accessed 9.5.2018): https://digitalcommons.wku.edu/ijesab/vol9/iss3/68/.

Oxford University Press, Oxford University Press Definition of "validation" in English in Oxford Dictonaries, web page. Available (accessed 9.5.2018): https://en.oxforddictionaries.com/definition/validation.

Paavolainen, L., Keijo Häkkinen, Ismo Hämäläinen, Nummela, A. & Rusko, H. (1999). Explosive-strength training improves 5-km running time by improving running economy and muscle power, Journal of applied physiology, Vol. 86(5), pp. 1527-1533. Available (accessed 10.8.2015): https://www.physiology.org/doi/full/10.1152/jappl.1999.86.5.1527.

Pan, J. & Tompkins, W.J. (1985). A Real-Time QRS Detection Algorithm, IEEE Transactions on Biomedical Engineering, Vol. 32(3), pp. 230-236.

Parak, J., Uuskoski, M., Machek, J. & Korhonen, I. (2017). Estimating Heart Rate, Energy Expenditure, and Physical Performance With a Wrist Photoplethysmographic Device During Running, JMIR Mhealth Uhealth, Vol. 5(7). Available (accessed 10.5.2018): https://mhealth.jmir.org/2017/7/e97/.

Parak, J. & Korhonen, I. Accuracy of Firstbeat Bodyguard 2 beat-to-beat heart rate monitor, Firstbeat Technologies Ltd., web page. Available (accessed 10.5.2018): https://assets.firstbeat.com/firstbeat/uploads/2015/11/white_paper_bodyguard2_final.pdf.

Parak, J., Tarniceriu, A., Renevey, P., Bertschi, M., Delgado-Gonzalo, R. & Korhonen, I. (2015). Evaluation of the beat-to-beat detection accuracy of PulseOn wearable optical heart rate monitor, 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Milan, 25 - 29 Aug., IEEE, pp. 8099-8102.

Patterson, J.A.C., McIlwraith, D.C. & Yang, G.Z. (2009). A Flexible, Low Noise Reflective PPG Sensor Platform for Ear-Worn Heart Rate Monitoring, 2009 Sixth International Workshop on Wearable and Implantable Body Sensor Networks, Berkeley, 3 - 5 June, IEEE, pp. 286-291.

Patterson, J.A.C. & Yang, G.Z. (2012). Dual-Mode Additive Noise Rejection in Wearable Photoplethysmography, 2012 Ninth International Conference on Wearable and Implantable Body Sensor Networks, London, 9 - 12 May, IEEE, pp. 97-102.

Patterson, J.A.C. & Yang, G.Z. (2011). Ratiometric Artifact Reduction in Low Power Reflective Photoplethysmography, IEEE Transactions on Biomedical Circuits and Systems, Vol. 5(4), pp. 330-338.

Pelizzo, G., Guddo, A., Puglisi, A., De Silvestri, A., Comparato, C., Valenza, M., Bordonaro, E. & Calcaterra, V. (2018). Accuracy of a Wrist-Worn Heart Rate Sensing Device during Elective Pediatric Surgical Procedures, Children, Vol. 5(3). Available (accessed 10.5.2018): http://www.mdpi.com/2227-9067/5/3/38.

Peng, F., Zhang, Z., Gou, X., Liu, H. & Wang, W. (2014). Motion artifact removal from photoplethysmographic signals by combining temporally constrained independent component analysis and adaptive filter, BioMedical Engineering OnLine, Vol. 13(1).

Available (accessed 10.5.2018): https://biomedical-engineering-online.biomedcentral.com/articles/10.1186/1475-925X-13-50.

Pietilä, J., Mehrang, S., Tolonen, J., Helander, E., Jimison, H., Pavel, M. & Korhonen, I. (2018). Evaluation of the accuracy and reliability for photoplethysmography based heart rate and beat-to-beat detection during daily activities, 2017 European Medical and Biological Engineering Conference (EMBEC), Tampere, 11 - 15 June, Springer, Singapore, pp. 145-148.

Pietilä, J., Helander, E., Myllymäki, T., Korhonen, I., Jimison, H. & Pavel, M. (2015). Exploratory analysis of associations between individual lifestyles and heart rate variability - based recovery during sleep, 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Milan, 25 - 29 Aug., IEEE, pp. 2339-2342.

Polar Electro Oy, RS800CX and H10 Heart Rate Sensor, web page. Available (accessed 10.5.2018): https://www.polar.com/en/products/.

Porto, L.G. & Junqueira, L.F.,Jr (2009). Comparison of time-domain short-term heart interval variability analysis using a wrist-worn heart rate monitor and the conventional electrocardiogram, Pacing and clinical electrophysiology : PACE, Vol. 32(1), pp. 43-51.

Pulse Oximeters - Premarket Notification Submissions [510(k)s] (2013). Pulse Oximeters - Premarket Notification Submissions [510(k)s] Guidance for Industry and Food and Drug Administration Staf, Food and Drug Administration, Rockville, 16 p.

PulseOn Oy, PulseOn Optical Heart Rate Monitor, web page. Available (accessed 10.5.2018): http://pulseon.com/.

R. Rox Anderson & John A. Parrish, (1981). The Optics of Human Skin, Journal of Investigative Dermatology, Vol. 77(1), pp. 13-19.

Reis, V.M., den Tillaar, R.V. & Marques, M.C. (2011). Higher Precision of Heart Rate Compared with VO2 to Predict Exercise Intensity in Endurance-Trained Runners, Journal of sports science & medicine, Vol. 10(1), pp. 164-168.

Reisner, A., Shaltis, P.A., McCombie, D. & Asada, H.H. (2008). Utility of the photoplethysmogram in circulatory monitoring, Anesthesiology, Vol. 108(5), pp. 950-958.

Renevey, P., Solà, J., Theurillat, P., Bertschi, M., Krauss, J., Andries, D. & Sartori, C. (2013). Validation of a wrist monitor for accurate estimation of RR intervals during sleep, 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Osaka, 3 - 7 July, IEEE, pp. 5493-5496.

Renevey, P., Vetter, R., Krauss, J., Celka, P. & Depeursinge, Y. (2001). Wrist-located pulse detection using IR signals, activity and nonlinear artifact cancellation, 2001 Conference Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Istanbul, 25 - 28 Oct., IEEE, pp. 3030-3033.

Robertson, A.H., King, K., Ritchie, S.D., Gauthier, A.P., Laurence, M. & Dorman, S.C. (2015). Validating the Use of Heart Rate Variability for Estimating Energy Expenditure, International Journal of Human Movement and Sports Sciences, Vol. 3(2), pp. 19-26.

Rousset, S., Fardet, A., Lacomme, P., Normand, S., Montaurier, C., Boirie, Y. & Morio, B. (2015). Comparison of total energy expenditure assessed by two devices in controlled and free-living conditions, European journal of sport science, Vol. 15(5), pp. 391-399.

Saalasti, S. (2003). Neural networks for heart rate time series analysis, University of Jyväskylä, 194 p. Available: https://jyx.jyu.fi/dspace/handle/123456789/13267.

Saalasti, S., Seppänen, M. & Kuusela, A. (2004). Artefact correction for heart beat interval data, Advanced Methods for Processing Bioelectrical Signals, Jyväskylä, October 2004, ProBisi Meeting, pp. 1-10.

Schäfer, A. & Vagedes, J. (2013). How accurate is pulse rate variability as an estimate of heart rate variability?: A review on studies comparing photoplethysmographic technology with an electrocardiogram, International journal of cardiology, Vol. 166(1), pp. 15-29.

Shaffer, F. & Ginsberg, J.P. (2017). An Overview of Heart Rate Variability Metrics and Norms, Frontiers in Public Health, Vol. 5(3). Available (accessed 10.5.2018): https://www.frontiersin.org/articles/10.3389/fpubh.2017.00258/full.

Shcherbina, A., Mattsson, C.M., Waggott, D., Salisbury, H., Christle, J.W., Hastie, T., Wheeler, M.T. & Ashley, E.A. (2017). Accuracy in Wrist-Worn, Sensor-Based Measurements of Heart Rate and Energy Expenditure in a Diverse Cohort, Journal of

personalized medicine, Vol. 7(2). Available (accessed 10.5.2018): http://www.mdpi.com/2075-4426/7/2/3.

Shi, P. (2009). Photoplethysmography in noninvasive cardiovascular assessment, Loughborough University, 152 p. Available: https://dspace.lboro.ac.uk/2134/5399.

Shi, P., Hu, S., Zhu, Y., Zheng, J., Qiu, Y. & Cheang, P.Y.S. (2009). Insight into the dicrotic notch in photoplethysmographic pulses from the finger tip of young adults, Journal of medical engineering & technology, Vol. 33(8), pp. 628-633.

Spierer, D.K., Rosen, Z., Litman, L.L. & Fujii, K. (2015). Validation of photoplethysmography as a method to detect heart rate during rest and exercise, Journal of medical engineering & technology, Vol. 39(5), pp. 264-271.

Spigulis, J., Gailite, L., Lihachev, A. & Erts, R. (2007). Simultaneous recording of skin blood pulsations at different vascular depths by multiwavelength photoplethysmography, Applied Optics, Vol. 46(10), pp. 1754-1759.

Stahl, S.E., An, H., Dinkel, D.M., Noble, J.M. & Lee, J. (2016). How accurate are the wrist-based heart rate monitors during walking and running activities? Are they accurate enough? BMJ Open Sport & Exercise Medicine, Vol. 2(1). Available (accessed 10.5.2018): https://bmjopensem.bmj.com/content/2/1/e000106.

Sztajzel, J. (2004). Heart rate variability: a noninvasive electrocardiographic method to measure the autonomic nervous system, Swiss medical weekly, Vol. 134(35-36), pp. 514-522.

Tanaka, H., Monahan, K.D. & Seals, D.R. (2001). Age-predicted maximal heart rate revisited, Journal of the American College of Cardiology, Vol. 37(1), pp. 153-156.

Tarniceriu, A., Parak, J., Renevey, P., Nurmi, M., Bertschi, M., Delgado-Gonzalo, R. & Korhonen, I. (2016). Towards 24/7 continuous heart rate monitoring, 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Orlando, 16 - 20 Aug., IEEE, pp. 186-189.

Tarvainen, M.P., Niskanen, J.P., Lipponen, J.A., Ranta-Aho, P.O. & Karjalainen, P.A. (2014). Kubios HRV — heart rate variability analysis software, Computer methods and programs in biomedicine, Vol. 113(1), pp. 210-220.

Teisala, T., Mutikainen, S., Tolvanen, A., Rottensteiner, M., Leskinen, T., Kaprio, J., Kolehmainen, M., Rusko, H. & Kujala, U.M. (2014). Associations of physical activity, fitness, and body composition with heart rate variability–based indicators of stress and recovery on workdays: a cross-sectional study, Journal of Occupational Medicine and Toxicology, Vol. 9(1). Available (accessed 10.5.2018): https://occup-med.biomedcentral.com/articles/10.1186/1745-6673-9-16.

Temko, A. (2015). Estimation of heart rate from photoplethysmography during physical exercise using Wiener filtering and the phase vocoder, 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Milan, 25 - 29 Aug., IEEE, pp. 1500-1503.

Terbizan, D.J., Dolezal, B.A. & Albano, C. (2002). Validity of Seven Commercially Available Heart Rate Monitors, Measurement in Physical Education and Exercise Science, Vol. 6(4), pp. 243-247.

The Qt Company, Qt Mobility, web page. Available (accessed 10.5.2018): https://github.com/qtproject/qt-mobility.

Tobaldini, E., Nobili, L., Strada, S., Casali, K.R., Braghiroli, A. & Montano, N. (2013). Heart rate variability in normal and pathological sleep, Frontiers in physiology, Vol. 4. Available (accessed 10.5.2018): https://www.frontiersin.org/articles/10.3389/fphys.2013.00294/full.

Tortora, G.J. & Grabowski, S.R. (2003). Principles of anatomy and physiology, 10th ed. Wiley, New York (NY), 1104 p.

TÜV SÜD, Certification mark for Fitness Trackers, web page. Available (accessed 9.5.2018): https://www.tuev-sued.de/product-testing/certificates/certification-mark-for-fitness-trackers.

Valenti, G. & Westerterp, K.R. (2013). Optical heart rate monitoring module validation study, 2013 IEEE International Conference on Consumer Electronics (ICCE), Las Vegas, 11 - 14 Jan., IEEE, pp. 195-196.

Vander, A.J., Sherman, J.H. & Luciano, D.S. (1990). Human physiology: the mechanisms of body function, 5th ed ed. McGraw-Hill, New York (NY), 724 p.

Vanderlei, L.C., Silva, R.A., Pastre, C.M., Azevedo, F.M. & Godoy, M.F. (2008). Comparison of the Polar S810i monitor and the ECG for the analysis of heart rate

variability in the time and frequency domains, Brazilian journal of medical and biological research, Vol. 41(10), pp. 854-859.

Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm, IEEE Transactions on Information Theory, Vol. 13(2), pp. 260-269.

Wallen, M.P., Gomersall, S.R., Keating, S.E., Wisloff, U. & Coombes, J.S. (2016). Accuracy of Heart Rate Watches: Implications for Weight Management, PloS ONE, Vol. 11(5). Available (accessed 10.5.2018): https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0154420.

Wang, R., Blackburn, G., Desai, M., Phelan, D., Gillinov, L., Houghtaling, P. & Gillinov, M. (2017). Accuracy of Wrist-Worn Heart Rate Monitors, JAMA cardiology, Vol. 2(1), pp. 104-106.

Webster, J.G. (1997). Design of pulse oximeters, Institute of Physics Publishing, Bristol, Philadelphia, 244 p.

Weippert, M., Kumar, M., Kreuzfeld, S., Arndt, D., Rieger, A. & Stoll, R. (2010). Comparison of three mobile devices for measuring R-R intervals and heart rate variability: Polar S810i, Suunto t6 and an ambulatory ECG system, European journal of applied physiology, Vol. 109(4), pp. 779-786.

Winokur, E.S., O'Dwyer, T. & Sodini, C.G. (2015). A Low-Power, Dual-Wavelength Photoplethysmogram (PPG) SoC With Static and Time-Varying Interferer Removal, IEEE Transactions on Biomedical Circuits and Systems, Vol. 9(4), pp. 581-589.

Winter, E.M., Jones, A.M. & Richard Davison, R.C. (ed.). 2006. Sport and Exercise Physiology Testing Guidelines. Taylor & Francis e-Library. 267 p.

World Medical Association (2013). World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects, JAMA, Vol. 310(20), pp. 2191-2194.

Wowern, E.v., Östling, G., Nilsson, P.M. & Olofsson, P. (2015). Digital Photoplethysmography for Assessment of Arterial Stiffness: Repeatability and Comparison with Applanation Tonometry, PLoS ONE, Vol. 10(8). Available (accessed 10.5.2018): http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4546304.

PUBLICATION I

EVALUATION OF WEARABLE CONSUMER HEART RATE MONITORS BASED ON PHOTOPLETYSMOGRAPHY

by

Parak, J. & Korhonen, I. (2014)

36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 3670-3673

Evaluation of Wearable Consumer Heart Rate Monitors Based on Photopletysmography

Jakub Parak, *IEEE Student Member* and Ilkka Korhonen, *IEEE Senior Member*

***Abstract*— Wearable monitoring of heart rate (HR) during physical activity and exercising allows real time control of exercise intensity and training effect. Recently, technologies based on pulse plethysmography (PPG) have become available for personal health management for consumers. However, the accuracy of these monitors is poorly known which limits their application. In this study, we evaluated accuracy of two PPG based (wrist i.e. Mio Alpha vs forearm i.e. Schosche Rhythm) commercially available HR monitors during exercise. 21 healthy volunteers (15 male and 6 female) completed an exercise protocol which included sitting, lying, walking, running, cycling, and some daily activities involving hand movements. HR estimation was compared against values from the reference electrocardiogram (ECG) signal. The heart rate estimation reliability scores for <5% accuracy against reference were following: mio Alpha 77,83% and Scosche Rhytm 76,29%. The estimated results indicate that performance of devices depends on various parameters, including specified activity, sensor type and device placement.**

I. Introduction

Heart rate monitoring is useful in wide areas including clinical medical care, pervasive health care, sports and well-being. HR describes an efficiency of cardiovascular system and heart functionality. People have been interested in HR monitoring since ancient Greek [1]. In 1960s Norman Holter invented a portable electrocardiogram (ECG) recorder and a HR analyzer [2]. Another milestone happened in 1982 when Polar Electro produced the first wearable HR monitor designed for sport purposes and based on ECG monitoring [3]. Today, ECG based HR monitors utilize usually chest strap and are widely available for consumers in affordable price. In parallel to chest strap based HR monitors, technologies based on photoplethysmogram (PPG) acquisition from wrist (REF), forearm (REF) or ear (REF) have been introduced. These solutions extend the use cases for HR monitoring by offering better comfort and more unobtrusive monitoring.

However, accuracy of these novel technologies has been little studied which limits their application especially beyond consumer use for recreational purposes. Chest strap based HR monitors, e.g. Polar Vantage XL, Polar Accurex, Cardioschamp and Cateye PL-6000, had a correlation >0.90 and standard error estimate <5 BPM during rest and moderate activity [4]. The best consumer level chest strap HR monitors have been found to provide comparable accuracy with ambulatory ECG in beat-to-beat detection and RR-interval estimation [5, 6]. Correlation coefficient of heart rate variability analysis demonstrated satisfactory correlation between Polar 810s and reference ECG during rest and ercocycling [7]. Another comparison study approves an interchangeability using of the Polar S810, Suunto t6 and ambulatory ECG system [8]. Smarthhealth watches and Polar Vintage XL were successfully validated against ambulatory ECG during four different loads on treadmill [9]. Comparison of the Actiheart and the Reynolds Holter system was performed in normal living conditions during common daily life activities [10]. In comparison, HR monitoring accuracy during treadmill running with finger photoplethymographs was found to be decreased as compared against the ECG reference [11]. A comparability problem of the Photopletysmography devices validation studies are discussed in the comprehensive expertise review [12].

The previous comparison studies are focused mainly on the traditional chest strap devices. Especially PPG based consumer targeted devices have not been objectively validated to date. In this study, we compare the accuracy of two different consumer wearable PPG based HR monitors during exercise against golden standard i.e. ECG based HR. We chose for comparison two different PPG based monitors (wrist and forearm worn devices). Materials and methods

A. Subjects

Twenty-one healthy volunteers (15 males and 6 females; 31,3 ± 10,7 years old) volunteered in the study. All participants were nonsmokers and they perform weekly some kind of physical activity. All subjects gave informed consent while participating the study.

B. Methods

Table 1 contains detailed description of protocol tasks and duration. Total testing time was 50 minutes. Selected protocol tasks focus to simulate intensive exercise, rest positions including sitting, lying on the bed in different positions and standing. Hand movements which can have significant impact of results were simulated in Rubic cube game play.

J. Parak is with Department of Signal Processing, Tampere University of Technology, Tampere, Finland (corresponding author e-mail: jakub.parak@tut.fi). He is also with Department of Circuit Theory, Faculty of Electrical Engineering, Czech Technical University in Prague, Prague, Czech Republic.

I. Korhonen is with Department of Signal Processing, Tampere University of Technology, Tampere, Finland (e-mail: ilkka.korhonen@tut.fi). He is also with VTT Technical Research Centre of Finland, Tampere, Finland.

TABLE I. TESTING PROTOCOL TASKS AND DURATION

Activity	*Duration [min]*
Rest sitting	4:00
Lying on bed on different positions	6:00
Standing	1:00
Walking 3km/h - 0% inclination	3:00
Walking 3km/h - 5% inclination	3:00
Walking 3km/h - 10% inclination	3:00
Walking 5km/h - 0% inclination	3:00
Walking 5km/h - 5% inclination	3:00
Walking 5km/h - 10% inclination	3:00
Running 9km/h - 0% inclination	3:00
Running 11km/h - 0% inclination	3:00
Rest sitting	2:00
Rest sitting and playing with Rubic cube	2:00
Rest sitting	2:00
Cycling 60 rpm	3:00
Cycling 90 rpm	3:00
Rest sitting	4:00

C. Data acquisition

HR was acquired with two PPG based HR monitors: Mio Alpha (Mio Global, Canada) and Schosche myRhyhm (Schosche Industries, CA, USA) (Figure 1).

Mio Alpha is worn on wrist and uses green LEDs and a photodetector for signal acquisition. Data were transmitted from device using the ANT+ technology to Garmin Forerunner device. HR data with timestamps were extracted from Garmin device for further analysis. Scosche Rhythm is worn on forearm and uses infrared LED and a photodetector for PPG acquisition. Data were transmitted by Bluetooth technology to iCardio Smartphone application where it was exported for further analysis. Both of devices were attached on subject body according manufacturers' recommendations.

The Embla Titanium multi-parameter wearable recorder was used for measuring the reference ECG signal. This device is designed for acquiring several biosignals including the ECG. Two ECG leads were acquired for reference heart rate estimation. Disposable electrodes were placed according two channels Holter measurement. [13]. Fixing of the disposable electrodes and cables were done by the medical tape for decreasing level of possible motion and other signal artifacts.

C. Statistical analyzes

Analysis of the reference ECG signal was performed with the Kubios HRV tool [14]. The better ECG RAW signal quality channel was selected by visual inspection of both recorded channels. The R-peaks were detected in selected channel by automatic R-peak detection algorithm which is included in HRV tool. In R-peak detection algorithm, QRS complexes are re-sampled at 2048 Hz with sinc-interpolation prior to R-peak detection to reduce the quantization error caused by low ECG sampling rate [15]. The all R-peak detections were verified manually in the reference signal. Heart timing signals algorithm was used for detection of the arrhythmias (ectopic beats) [16]. These beat were excluded from the final statistical evaluation and error estimation.

The evaluated and reference heart rate signals were resampled to 10 Hz sampling frequency. HR acquired from PPG HR monitors and reference HR were synchronized in time by applying cross-correlation function between the reference and the target HR and by maximizing the cross-correlation value at t=0. The signals were smoothed by moving average in 5s second window.

Several HR detection accuracy parameters were evaluated for both of tested device. The successful HR score for < 5% and <10% beats per minutes difference against reference were calculated in 5s average HR window without overlaps. Mean error (ME), mean absolute error (MAE), mean percentage error (MPE), mean absolute percentage error (MAPE) between the mio Alpha, Scosche Rhythm and reference HR were calculated.

a)

b)

c)

Figure 1. a) Mio Alpha. b) Scosche myRhythm c) Test configuration

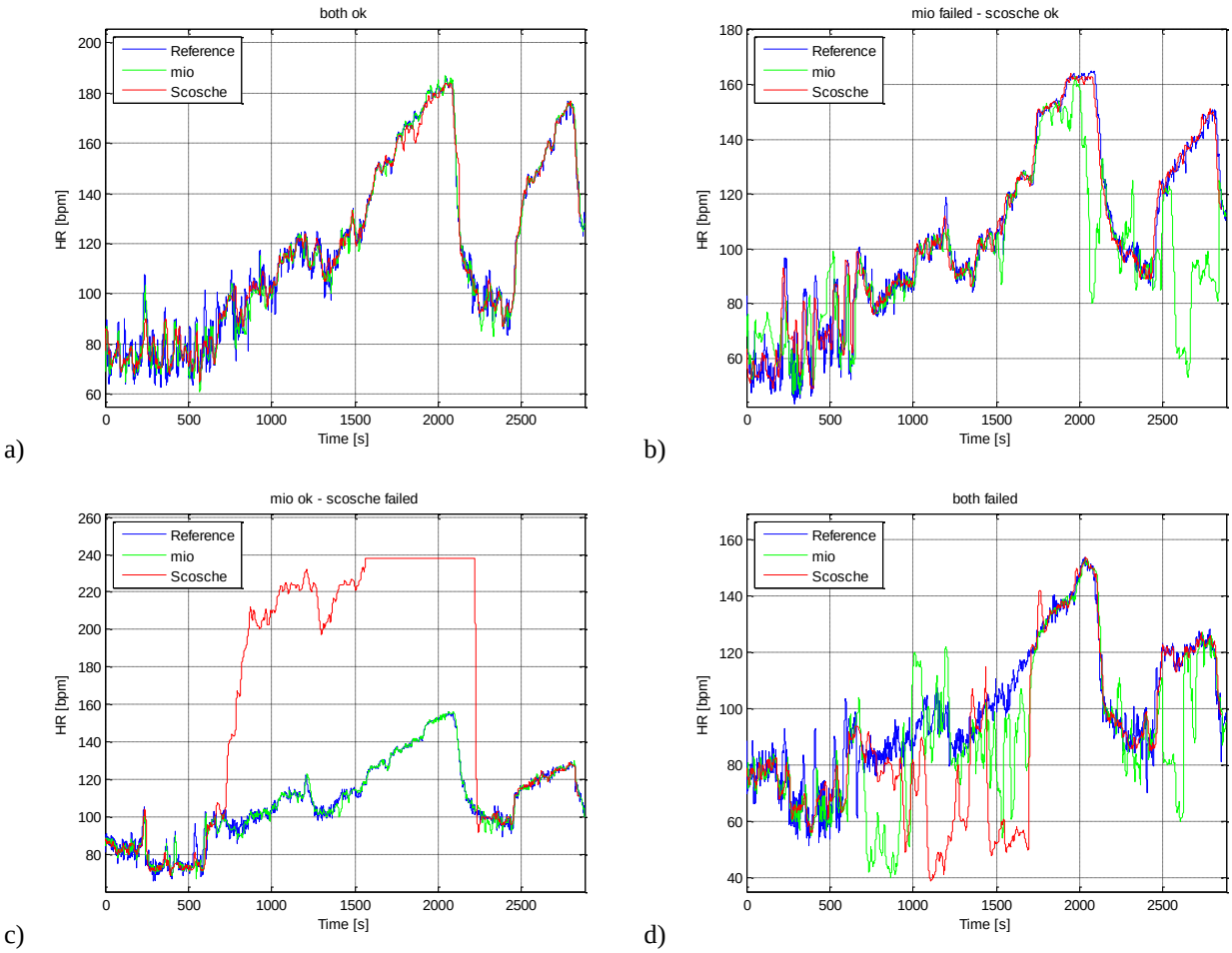


Figure 2. A) HR monitoring successful with both devices. B) Wrist based Mio Alpha fails during ergocycling. C) Forearm based Schosche Rhythm failed during walking and running, likely due to sensor displacement. D) Both devices show poor performance.

TABLE II. MIO ALPHA ERROR STATISTICS AS COMPARED TO REFERENCE ECG BASED HR (N=21)

Activity	***Mean Error* [bpm]**	***Mean Error* [%]**	***Mean Abs Error* [bpm]**	***Mean Abs Error* [%]**
global	-1,21	-1,74	4,43	5,23
rest	-0,20	-0,52	3,92	5,37
walking	-0,89	-1,72	4,98	5,60
running	-2,26	-1,93	2,89	2,37
cycling	-3,76	-4,80	4,64	5,53
rubic	-1,26	-1,83	7,54	8,43

TABLE III. SCOSCHE MYRHYTHM ERROR STATISTICS AS COMPARED TO REFERENCE ECG BASED HR (N=21)

Activity	***Mean Error* [bpm]**	***Mean Error* [%]**	***Mean Abs Error* [bpm]**	***Mean Abs Error* [%]**
global	1,11	-1,62	6,82	6,78
rest	0,07	-1,43	4,83	5,96
walking	1,83	-3,13	10,48	10,49
running	3,28	0,63	6,75	3,81
cycling	-0,89	-0,87	1,84	1,73
rubic	2,59	1,46	4,73	3,94

I. Results

Both PPG HR monitors were able to monitor HR during exercise but not without errors in some cases. Representative examples are presented in Figure 2. HR estimation success rates for different activities are reported on Table II. Average performance was similar in both devices but Mio Alpha performed better during walking and running and Schosche Rhythm during cycling and Rubik's cube. Estimation error for Mio Alpha and Schosche Rhythm are presented in Tables III and IV, correspondingly.

TABLE II. MIO ALPHA AND SCHOSCHE RHYTHM SUCCESS RATES DURING DIFFERENT ACTIVITIES (N=21)

	Mio Alpha		***Scosche myRhythm***	
Activity	***score <5%***	***score <10%***	***score <5%***	***score <10%***
global	77,83	87,49	76,29	86,26
rest	72,53	84,87	69,53	83,88
walking	76,53	87,18	71,64	81,76
running	94,58	96,18	90,97	93,26
cycling	87,71	91,74	92,22	97,43
rubic	51,46	72,29	80,21	91,88

IV. Discussion

We evaluated new PPG based HR monitors against reference (ECG) HR. The results show that the PPG based HR monitors are able to monitor HR during exercise but not without errors. On average, PPG based HR was within 10bpm from true HR 86-87% of the time. This may be considered as satisfactory overall performance. However, sometimes the monitors fail to monitor HR and in such cases grand errors are seen (see Fig 1).

Wrist based monitor (Mio Alpha) performed better during walking and running while forearm based Scosche Rhythm was better during cycling and hand movements (Rubik's cube). It is natural that forearm based sensor is less affected by hand movements which certainly occur in Rubik's cube test but likely also during cycling, related to using hands for balancing and holding the steering while cycling. Poorer performance of the forearm based device during running and walking is, however, slightly surprising as forearm should be objected to lower level of accelerations than wrist also during these activities. The difference may hence be related to different implementation issues, such as algorithms used to extract HR, or sensor arrangements (e.g. use of different wavelengths in PPG acquisition).

The average performance of both devices was satisfactory but momentary grand errors reduce the usefulness of them. Our data does not allow to study the exact reasons for failures. However, poor sensor placement or attachment, or displacement of the sensor during exercising, may explain some of the errors. If optical coupling between the sensor and the tissue is not maintained steady during the monitoring, the loss of signal and hence ability to monitor HR will result.

Our results demonstrate that new PPG based HR monitors are becoming a real option for consumer HR monitoring at least during exercising, when ultimate performance is not required. However, the PPG monitors studied in this paper do not yet reach the level of reliability of the chest strap based HR monitors. The reduced accuracy is partially compensated by better usability and comfort.

References

[1] G. E. Billman, "Heart rate variability - a historical perspective," *Frontiers in physiology*, vol. 86, pp. 1–13, Nov. 2011.

[2] N. Holter, "New methods for Heart Studies," *Science*, vol. 134, pp. 1214–1219, Oct. 1961.

[3] R.M. Laukkanen, P.K. Virtanen, "Heart rate monitors: state of the art," *Journal of Sport Sciences*, pp. 3 – 7, Jan. 1998.

[4] D. J. Terbizan, B. A. Dolezal and Ch. Albano, "Validity of Seven Commercially Available Heart Rate Monitors," *Measurement in Physical Education and Exercise Science*, vol. 6, pp. 243–247, 2002.

[5] L. C. Vanderlei, R. A. Silva, C. M .Pastre, F. M. Azevedo and F. M. Godoy, "Comparison of the Polar S810i monitor and the ECG for the analysis of heart rate variability in the time and frequency domains." *Brazilian Journal of Medical and Biological Research*, vol. 41, pp. 854 – 859, Oct. 2008

[6] L. G. Porto, L. F. Jr. Junqueira, Comparison of Time-Domain Short-Term Heart Interval Variability Analysis Using a Wrist-Worn Heart Rate Monitor and the Conventional Electrocardiogram, " *Pacing and Clinical Electrophysiology*, vol. 32, pp. 43 – 51, Jan. 2009

[7] M. Kingsley, M. J. Lewis and R. E. Marson, "Comparison of Polar 810 s and an Ambulatory ECG System", *International Journal of Sports Medicine*, vol. 26 ,pp. 39 – 43, Jan. – Feb,

[8] M. Weippert, M. Kumar, S. Kreuzfeld, D. Arndt, A. Rieger and R. Stoll, "Comparison of three mobile devices for measuring R–R intervals and heart rate variability: Polar S810i, Suunto t6 and an ambulatory ECG system," *European Journal of Applied Physiology*, vol. 109, pp. 779 – 786, Jul. 2010.

[9] C. M. Lee, M. Gorelick, "Validity of the Smarthealth Watch to Measure Heart Rate During Rest and Exercise", *Measurement in Physical Education and Exercise Science*, vol. 15, pp. 18 – 25, Jan. 2011.

[10] J. Kristiansen, M. Korshøj. J. H. Skotte, T. Jespersen, K. Søgaard, K. Mortensen and A. Holtermann, "Comparison of two systems for long-term heart rate variability monitoring in free-living conditions - a pilot study." *BioMedical Engineering OnLine*, pp. 10 – 27, Apr. 2011.

[11] G. B. David. J. Araujo and T. R. Thomas, "A procedure for evaluating portable heart monitors," *Behavior Research Methods, Instruments, & Computers*, vol. 16, pp. 7 – 11, Jan. 1984.

[12] A. Schäfer, J. Vagedes, "How accurate is pulse rate variability as an estimate of heart rate variability? A review on studies comparing photoplethysmographic technology with an electrocardiogram", *International Journal of Cardiology*, vol. 166, pp. 15 – 29, Jun 2013.

[13] Schiller (2011). Schiller electrode placement for Holter MT-101 (and MT-101 nano) with 2-channel recording [Online cited 2014 April 7]. Available: http://www.youtube.com/watch?v=Mwb2ZffQEtk

[14] Biosignal Analysis and Medical Imaging Group (2012). Kubios HRV - Heart Rate Variability Analysis Software [Online cited 2014 April 7]. Available: http://kubios.uef.fi/KubiosHRV/

[15] M. P. Tarvainen, J. P. Niskanen, J. A. Lipponen, P. O. Ranta-Aho, and P. A Karjalainen., "Kubios HRV - Heart rate variability analysis software," *Computer Methods and Programs in Biomedicine*, vol. 113, pp. 210 – 220, Jan. 2014.

[16] J. Mateo, P. Laguna P, "Analysis of heart rate variability in the presence of ectopic beats using the heart timing signal," *IEEE Transactions on Biomedical Engineering*, vol. 50, pp. 334 - 343, Mar. 2003.

PUBLICATION II

EVALUATION OF ACCURACY AND RELIABILITY OF PULSEON OPTICAL HEART RATE MONITORING DEVICE

by

Delgado-Gonzalo, R., Parak, J., Tarniceriu, A., Renevey, P.,
Bertschi, M. & Korhonen, I. (2015)

37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 430-433

Evaluation of Accuracy and Reliability of PulseOn Optical Heart Rate Monitoring Device

Ricard Delgado-Gonzalo, Jakub Parak, *IEEE Student Member*, Adrian Tarniceriu, Philippe Renevey, Mattia Bertschi, and Ilkka Korhonen, *IEEE Senior Member*

***Abstract*— PulseOn is a wrist-worn optical heart rate (HR) monitor based on photoplethysmography. It utilizes multi-wavelength technology and optimized sensor geometry to monitor blood flow at different depths of skin tissue, and it dynamically adapts to an optimal measurement depth in different conditions. Movement artefacts are reduced by adaptive movement-cancellation algorithms and optimized mechanics, which stabilize the sensor-to-skin contact. In this paper, we evaluated the accuracy and reliability of PulseOn technology against ECG-derived HR in laboratory conditions during a wide range of physical activities and also during outdoor sports. In addition, we compared the performance to another on-the-shelf consumer product Mio LINK®. The results showed PulseOn reliability (% of time with error <10bpm) of 94.5% with accuracy (100% - mean absolute percentage error) 96.6% as compared to ECG (vs 86.6% and 94.4% for Mio LINK®, correspondingly) during laboratory protocol. Similar or better reliability and accuracy was seen during normal outdoor sports activities. The results show that PulseOn provides reliability and accuracy similar to traditional chest strap ECG HR monitors during cardiovascular exercise.**

I. Introduction

Wearable monitoring of heart rate (HR) during physical activity and exercising allows real-time control of exercise intensity and training effect. Chest strap HR monitors based on electrocardiography (ECG) have been the standard for sports HR monitoring for 20 years. Chest strap based HR monitors typically have a correlation of >0.90 and a standard error estimate <5 BPM during rest and moderate activity, which is considered sufficient for consumer sports use [1]. The best chest strap HR monitors have been found to provide comparable accuracy with ambulatory ECG HR monitoring [2, 3, 4]. However, discomfort and complication of use has limited their popularity among consumers. Optical HR monitoring allows an unobtrusive and comfortable alternative for HR monitoring during exercise. However, most products up to today have suffered from poor reliability and accuracy [5]. In this paper, we evaluate a PulseOn optical HR monitor and evaluate it against ECG-based HR monitoring as well as against another on-the-shelf consumer optical HR monitor Mio LINK® in laboratory conditions. We also show that the system is robust to real-life outdoor conditions.

R. Delgado-Gonzalo, Ph. Renevey, and M. Bertschi are with CSEM - Centre Suisse d'Electronique et Microtechnique, Jacquet-Droz 1, 2002 Neuchâtel, Switzerland (corresponding e-mail : ricard.delgado@csem.ch).

J. Parak and I. Korhonen are with Department of Signal Processing, Tampere University of Technology, Tampere, Finland, and PulseOn Oy, Espoo, Finland.

A. Tarniceriu is with PulseOn SA, Neuchâtel, Switzerland.

II. Photoplethysmography and Wearable HR Monitoring

A. Physiological Principles

PulseOn is based on photoplethysmography where skin tissue is illuminated with a light source (typically LED) and the intensity of light that has propagated through the tissue is measured with a photodetector (PD) [6]. The blood volume in the small peripheral vessels close to the skin (see Figure 1) is varying with the pumping action of the heart and causes variations in the propagated light intensity. By analyzing these variations it is possible to derive HR. However, there are several factors which affect the light propagation and hence make reliable optical HR monitoring highly challenging.

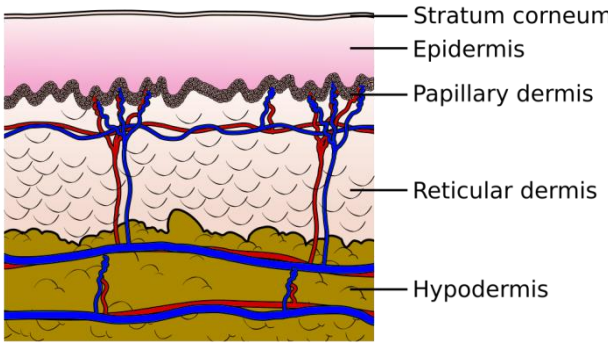


Figure 1. Structure of human skin. The thickness of papillary dermis and reticular dermis, where blood flow mainly occurs, vary between 0.6 and 3mm. When the skin is cold, perfusion in papillary dermis is minimal and is reduced also in reticular dermis.

First, the human skin is a complex non-homogeneous structure (Figure 1). Thus, even a small displacement of the sensor or a change in the sensor-skin contact may cause significant changes in the light propagation path [6]. This makes the technology very prone to movement artefacts. Furthermore, human physiology, especially temperature control, causes peripheral vascular dilatation and constriction depending on multiple factors (*e.g.*, environmental temperature, intra-body heat production), and hence both the volume of blood close to skin and the depth of main perfusion varies greatly between conditions and individuals. Finally, there are inter-individual differences in the skin and tissue structures and thickness of the layers as well as the amount of melanin. As a result, the optimal measurement depth as well as the strength of the signal varies greatly between situations and individuals. Measurement depth may be optimized by the selection of the light color (wavelength) and design of the sensor layout [6]. The depth of penetration

of the light into the tissue depends on light wavelength and distance between the LED and the PD [6]. Due to optical properties of the different components of the tissue, longer wavelengths, such as infra-red light (IR), will penetrate deeper into the tissue than short wavelengths (*e.g.*, green light). In addition, shortening the PD-LED distance will reduce the average light propagation path, and vice versa. Green light with short PD-LED distance is able to illuminate blood flow only very close to skin while IR light and longer PD-LED distance provide deeper measurement. However, the larger the measurement area is, more prone to movement artefacts the measurement is [6]. As a result, green light and short PD-LED distance are less sensitive to movement artefacts [7] but more sensitive to poor perfusion (*e.g.*, during cold skin). Typical optical HR solutions compromise between these demands and provide measurement on a single wavelength and single PD-LED distance, resulting in a single average measurement depth in all conditions and individuals, making it sensitive to variations in blood perfusion (*e.g.*, cold skin), individual differences, and/or movements.

Optomechanical design of the sensing device affects the signal quality significantly [6]. Reducing the weight of the sensing device, reduces forces caused by the movement and thereafter the skin-sensor pressure changes during exercise

B. PulseOn Technology

PulseOn sensor solution[1] takes advantage of multiple light wavelengths and optimally matched LED-PD distances to allow the measurement of blood flow in different tissue depths (see Figure 2). It dynamically chooses the optimal combination for reliable and accurate HR monitoring. The use of green light and short PD-LED distance allows for a robust HR monitoring, even during intense movements, while the use of IR and longer PD-LED distance allows for HR acquisition even during low blood perfusion (*e.g.*, cold skin).

Figure 2. PulseOn sensor solution combines green and IR light with optimally matched LED-PD distances.

The mechanical design[2] of the housing and the strap provides PulseOn with a stable skin-sensor contact in a wide range of conditions without compromising the comfortable use. The design also reduces artefacts and improves HR reliability. The light weight (29g, including strap) further reduces artefacts and improves usage comfort. Intelligent algorithms analyze PD signals and decide the optimal measurement combination in each situation. HR detection algorithm applies integrated accelerometer data to reduce movement artefacts and provides accurate HR estimation even during very intensive training, spanning up to full running speeds and maximum HR levels.

III. EXPERIMENTAL VALIDATION

A. Controlled Laboratory Protocol

The test group consisted on N=19 healthy volunteers, from which 9 are men and 10 women (see Table I). All participants were nonsmokers and physically active[3].

TABLE I. SUBJECTS' ANTHROPOMETRIC PARAMETERS

Characteristic	μ ± σ	Range
Age (years)	28.30 ± 5.69	23 – 47
Height (m)	1.74 ± 0.11	1.55 – 1.90
Weight (kg)	72.30 ± 12.59	52 – 99

The subjects followed a standardized protocol that included a wide set of activities, ranging from sedentary to vigorous and causing rapid and wide variations in HR, recorded in laboratory settings (see Table II). The treadmill and ergocycle used in the execution of the protocol were Daum Ergo Run Premium Alpha 24 and the Tunturi Alpha 300 respectively.

TABLE II. TESTING LABORATORY PROTOCOL AND DURATIONS

Activity	Duration
Standing	1min
Walking on a treadmill at 3km/h, 0% inclination	3min
Walking on a treadmill at 3km/h, 5% inclination	3min
Walking on a treadmill at 3km/h, 10% inclination	3min
Walking on a treadmill at 5km/h, 0% inclination	3min
Walking on a treadmill at 5km/h, 5% inclination	3min
Walking on a treadmill at 5km/h, 10% inclination	3min
Running on a treadmill at 9km/h, 0% inclination	3min
Running on a treadmill at 11km/h, 0% inclination	3min
Rest sitting	4min
Cycling 60rpm*	3min
Cycling 90rpm*	3min
Rest sitting	4min

*Unconditioned males (activity class <5): 50 Watts
*Unconditioned females (activity class <5): 50Watts
*Conditioned males (activity class 5 or above): 100Watts
*Conditioned females (activity class 5 or above): 75Watts

HR signals were acquired with Mio LINK® though a Garmin Forerunner 610 (ANT device) and the PulseOn's HR monitor. The chest-strap ECG Polar Electro RS800CX HR monitor was used as the reference. This chest strap provides an ECG-level accuracy of the HR during sports [4]. PulseOn, Mio LINK®, and reference HR signals were synchronized in time by maximizing the cross-correlation among the signals. This process resulted in comparable and time-synced HR signals among all devices. Then, data was resampled to the same rate and averaged over 5s windows. PulseOn HR performance was estimated by the following parameters:

- *Reliability:* % of time that the absolute error is smaller than 10bpm.
- *Accuracy:* (100% - Mean Absolute Percentage Error).

[1] Patent pending.
[2] Patent pending.

[3] All participants gave informed consent to participate in the study. The study was conducted according to Helsinki Declaration.

The reliability provides a sense of the amount of time the system is working within an acceptable confidence interval, and the accuracy provides a sense of the error committed by the system at any point in time.

We show in Table III the mean performance indicators for specific activities (resting, walking, running, and biking) as well as global values for the whole protocol. PulseOn had significantly better global performance than Mio LINK® during the protocol (reliability 94.5% vs 86.6% and accuracy 96.6% vs 94.3% for PulseOn and Mio LINK®, correspondingly). The difference was mainly caused by the walking activity, where PulseOn reaches an average reliability of 90.8% and an average accuracy of 95.8% whereas Mio LINK® obtains the values of 73.7% and 90.2% respectively.

TABLE III. MEAN PERFORMANCE INDICATORS OF PULSEON AND MIO LINK® DURING THE LABORATORY PROTOCOL

Activity	PulseOn		Mio LINK®	
	Reliability (%)	Accuracy (%)	Reliability (%)	Accuracy (%)
Rest	97.9	97.1	97.4	97.3
Walking	90.8	95.8	73.7	90.2
Running	99.4	98.0	99.8	98.8
Cycling	96.0	96.8	97.0	97.7
Protocol	94.5	96.6	86.6	94.3

In Figure 3, we show the Bland-Altman plots comparing the error distributions of PulseOn and Mio LINK®. As expected from Table III, the error distributions for rest, running, and cycling are very similar for both devices. Walking is the activity that has greater dispersion on both devices. However, we can observe a larger dispersion for Mio LINK® around the interval [80,140] bpm. This has a clear impact on the global performance for the complete protocol.

B. Outdoors Testing

Subjects from Section III.A were randomly assigned to perform physical exercises outdoors. We recorded a total of 24 events that included track-running, trail-running, urban-running, walking, track-cycling, and road-cycling. Then, we grouped the recordings by their dominant activity in one of the following classes: Walking, Running, or Cycling.

We used the chest-strap ECG Polar Electro RS800CX HR monitor to obtain a reference heart rate. The data obtained from PulseOn's HR monitor was averaged over 5s windows and resampled in order to match the same sampling rate as the reference signal.

In Table IV, we show mean performance indicators of the PulseOn's HR monitor for each activity category. We can see that the values obtained, in terms of reliability and accuracy, are equivalent to those found in the controlled laboratory protocol of Section III.A. These values validate the PulseOn's HR monitor technology and show that the system is robust to real-life outdoor conditions (*e.g.*, changes in temperature, wind, non-uniform pace).

TABLE IV. MEAN PERFORMANCE INDICATORS OF PULSEON IN OUTDOOR ACTIVITIES

Main activity	PulseOn	
	Reliability (%)	Accuracy (%)
Walking (N=3)	94.1	96.6
Running (N=17)	99.1	97.9
Cycling (N=4)	95.2	97.3
Mean (N=24)	97.8	97.6

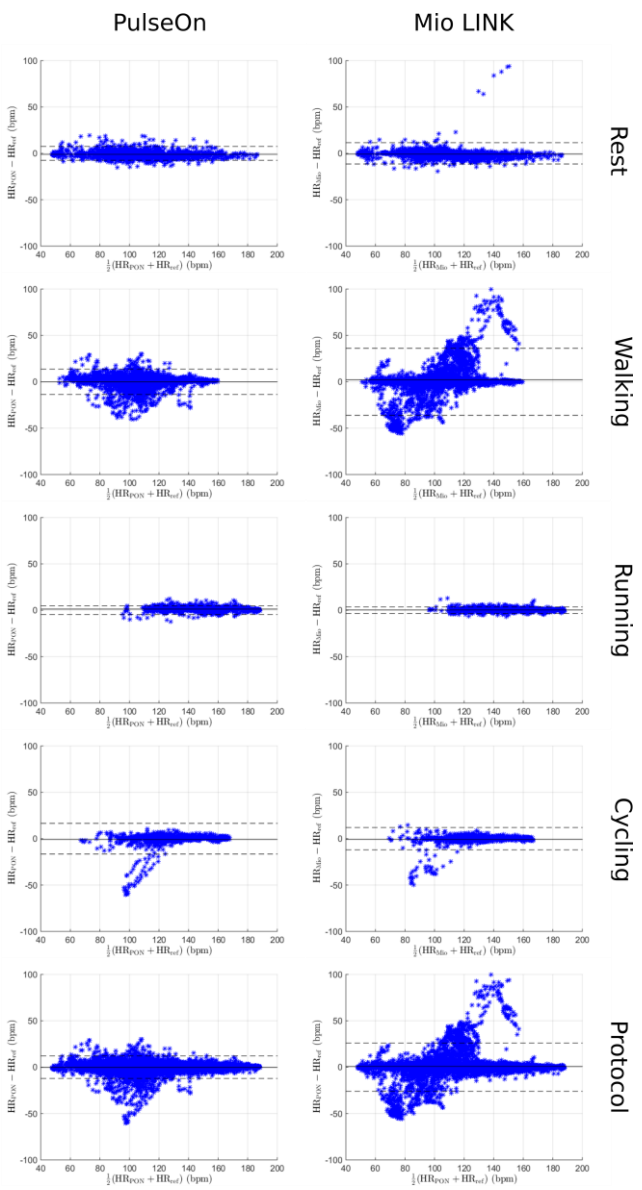


Figure 3. Bland-Altman plots comparing the reference ECG-derived HR to the HR derived from the wrist-devices. The left column compares the PulseOn monitor to the reference, and the right column compares the Mio LINK® to the reference. The rows from, top to bottom, correspond to the following activities: rest, walking, running, biking, and full protocol. HR_{PON}, HR_{Mio}, and HR_{ref} refer to the heart rate estimated with the PulseOn monitor, Mio LINK®, and the reference respectively.

In Figure 4, we compare the performance of the PulseOn monitor against the ECG-based reference for several outdoor activities. The PulseOn's HR monitor is capable of following the reference in stationary situation (*e.g.*, Figure 4B, C, G), as well as in fast changing heart rates (*e.g.*, Figure 4E, H). These differences are activity-dependent and reflect intrinsic characteristics from the type of exercise. For instance, Figure 4H shows the heart rate while driving a road bike in the city. The steep variations of the heart-rate values are linked to periods of time where the subject was steady in traffic lights.

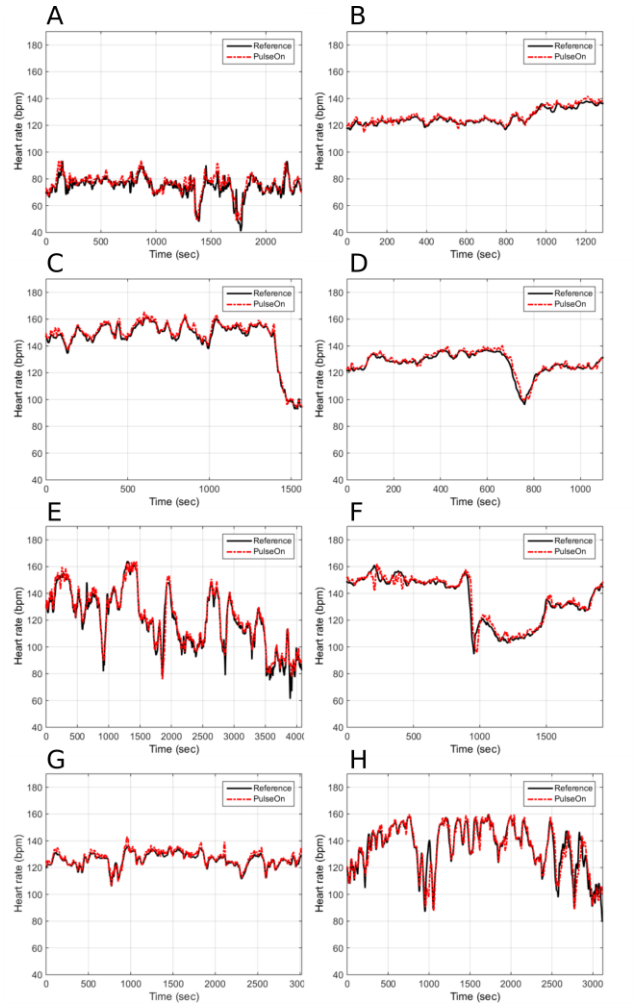


Figure 4. Comparison of the performance of the PulseOn monitor against the ECG-based reference, for several outdoor activities. A) Walking with two stops. B) Outdoor running. C) Outdoor running for 22 minutes and walking for 3 minutes. D) Outdoor running. E) Trail running in Lapland. F) Trail running in Lapland to Kiilopää. G) Outdoor cycling. H) Road-biking in traffic.

IV. Conclusion

PulseOn HR monitor measures blood flow in different depths of skin tissue and adapts to different situations and individual differences. PulseOn mechanical and sensor optimization provide reliability and accuracy comparable to ECG based chest belt HR monitors [1, 3, 2, 4] during typical cardiovascular exercises.

The results showed that PulseOn's mean reliability is 94.5% with an accuracy of 96.6%, opposed to 86.6% and 94.3% of Mio LINK®.

We also provided evidence of the robustness of the system in outdoor activities such as trail-running, urban-running, walking, track-cycling, and road-cycling. Under these conditions, PulseOn's HR monitor obtained an accuracy of 97.8% and a reliability of 97.6%.

V. References

[1] D. Terbizan, B. Dolezal and C. Albano, "Validity of seven commercially available heart rate monitors," *Measurement in Physical Education and Exercise Science*, vol. 6, pp. 243-247, 2002.

[2] L. Porto and L. Junqueira, "Comparison of time-domain short-term heart interval variability analysis using a wrist-worn heart rate monitor and the conventional electrocardiogram," *Pacing and Clinical Electrophysiology*, vol. 32, pp. 43-51, January 2009.

[3] L. Vanderlei, R. Silva, C. Pastre, F. Azevedo and F. Godoy, "Comparison of the Polar S810i monitor and the ECG for the analysis of heart rate variability in the time and frequency domains," *Brazilian Journal of Medical and Biological Research*, vol. 41, pp. 854-859, October 2008.

[4] M. Kingsley, M. Lewis and R. Marson, "Comparison of Polar 810s and an ambulatory ECG system," *International Journal of Sports Medicine*, vol. 26, pp. 39-43, 2005.

[5] J. Parak and I. Korhonen, "Evaluation of wearable consumer heart rate monitors based on photopletysmography," in *36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Chicago, IL, USA, 2014.

[6] M. Bertschi, P. Renevey, J. Solà, M. Lemay, J. Parak and I. Korhonen, "Application of optical heart rate monitoring," in *Wearable Sensors. Fundamentals, Implementation and Applications*, Elsevier Academic Press, 2014, p. 656.

[7] Y. Maeda, M. Sekine, T. Tamura, A. Moriya, T. Suzuki and K. Kameyama, "Comparison of reflection green light and infrared photoplethysmography," in *30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Vancouver, BC, Canada, 2008.

PUBLICATION III

EVALUATION OF THE BEAT-TO-BEAT DETECTION ACCURACY OF PULSEON WEARABLE OPTICAL HEART RATE MONITOR

by

Parak, J., Tarniceriu, A., Renevey, P., Bertschi, M.,
Delgado-Gonzalo, R. & Korhonen, I. (2015)

37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 8099-8102

Evaluation of the Beat-to-Beat Detection Accuracy of PulseOn Wearable Optical Heart Rate Monitor

Jakub Parak, *IEEE Student Member,* Adrian Tarniceriu, Philippe Renevey, Mattia Bertschi, Ricard Delgado-Gonzalo and Ilkka Korhonen, *IEEE Senior Member*

***Abstract*— Heart rate variability (HRV) provides significant information about the health status of an individual. Optical heart rate monitoring is a comfortable alternative to ECG based heart rate monitoring. However, most available optical heart rate monitoring devices do not supply beat-to-beat detection accuracy required by proper HRV analysis. We evaluate the beat-to-beat detection accuracy of a recent wrist-worn optical heart rate monitoring device, PulseOn (PO). Ten subjects (8 male and 2 female; 35.9±10.3 years old) participated in the study. HRV was recorded with PO and Firstbeat Bodyguard 2 (BG2) device, which was used as an ECG based reference. HRV was recorded during sleep. As compared to BG2, PO detected on average 99.57% of the heartbeats (0.43% of beats missed) and had 0.72% extra beat detection rate, with 5.94 ms mean absolute error (MAE) in beat-to-beat intervals (RRI) as compared to the ECG based RRI BG2. Mean RMSSD difference between PO and BG2 derived HRV was 3.1 ms. Therefore, PO provides an accurate method for long term HRV monitoring during sleep.**

I. Introduction

New wearable sensing technologies provide unobtrusive, comfortable and affordable methods for long-term real life monitoring of health and physiological status of the users. Multiple commercial devices have been recently released allowing measurement of e.g. physical activity, heart rate and sleep. However, there is an increasing need to evaluate the accuracy of these new technologies and compare them to established gold standards.

Heart rate variability (HRV) provides significant information about the health status of an individual. HRV may be used in a wide spectrum of applications, such as clinical practice [1], sleep quality measurement [2], and stress and recovery analysis [3]. Accurate detection of beat-to-beat heart rate is necessary for the analysis of the HRV [4, 5]

J. Parak and I. Korhonen are with the Department of Signal Processing, Tampere University of Technology, Tampere, Finland, and PulseOn Oy, Espoo, Finland. (corresponding e-mail : jakub.parak@tut.fi).

A. Tarniceriu is with PulseOn SA, Jacquet-Droz 1, 2002, Neuchâtel, Switzerland.

R. Delgado-Gonzalo, Ph Renevey, and M. Bertschi are with CSEM - Centre Suisse d'Electronique et Microtechnique, Jacquet-Droz 1, 2002 Neuchâtel, Switzerland .

Several studies have been evaluated chest strap (ECG) based wearable heart rate monitors for HRV detection accuracy [6, 7, 8, 9]. These devices were usually compared against ambulatory ECG recorders in controlled laboratory conditions. Published results present high accuracy for the estimation of beat-to-beat (RR) intervals using chest strap devices, with limits of agreements for group differences of less than ±10ms. On the other hand, wearing chest straps is uncomfortable in long term heart rate monitoring applications, especially during sleep. In addition, dry skin and poor skin contact often disturb chest strap based HRV monitoring during sleep.

Photoplethysmography (PPG) provides an alternative method to monitor HRV [10, 11,12]. In [13, 14] the accuracy of HRV extraction from PPG during sleep was compared against ECG based RR Holter recordings. All of the mentioned studies used wearable devices based on infrared and red LED reflective PPG sensing technology, with encouraging results. However, it has been suggested that pulse rate variability (PRV) from PPG is sufficiently accurate only for healthy (and mostly younger) subjects at rest, HRV estimated from PRV tends to be overestimated against ECG based values, and motion artifacts lead to inaccurate PPG-based beat detection [15].

The aim of this study is to evaluate the accuracy of the beat-to-beat detection of the PulseOn (PO) consumer wearable optical heart rate monitor. The comparison is performed against the Firstbeat Bodyguard 2 (BG2) wearable RR interval recorder. The application of Firstbeat beat–to–beat data artifact correction algorithm, estimation of HRV parameters and energy expenditure are also presented in this paper.

II. Methods

A. Subjects

Ten healthy volunteers (8 male and 2 female; 35.9±10.3 years old) participated in this study. Two male participants were ex-smokers. All subjects perform moderate physical training weekly. A total of 13 recordings were obtained for this study.

The experimental procedures described in this paper complied with the principles of Helsinki Declaration of 1975, as revised in 2000. All subjects gave informed consent to

participate and they had a right to withdraw from the study at any time. Their information was anonymized prior the analysis.

B. Data recording conditions

Subjects performed the recordings at their homes in normal bedroom sleeping conditions. The average non-stop recorded sleep time of all subjects was 5.1±1.2 hours. The recordings did not span the whole nights due to battery limitations. The subjects were instructed to start the recordings as soon as they went to bed.

C. Data acquisition

PO (PulseOn Technologies Ltd, Espoo, Finland, www.pulseon.fi , [accessed 31.03.2015]) is a wearable wristband consumer optical heart rate monitor with double wavelength technology (green and infrared) and optimized optical sensors for high accuracy signal measurements (see Figure 1) [14]. The device was worn as instructed by the manufacturer on non-dominant hand, about one finger width from the wrist bone, and tightened by the subjects so that the skin contact was firm but still comfortable for the whole night recording. Beat-to-beat HR was detected automatically by the device. Data was logged to PO mobile phone application and uploaded for further processing offline.

Figure 1. PulseOn consumer wearable optical based heart rate monitor with double wavelenght optical sesing technology

The reference RR intervals were acquired with BG2 (Firstbeat Technologies Ltd, Jyväskyla, Finland, www.firstbeat.fi , [accessed 31.03.2015]) long term ECG based recorder with two disposable electrodes (see Figure 2). This device provides standard RR interval precision in milliseconds. The beat-to-beat detection accuracy and artifact correction algorithm of the selected reference device was evaluated in a laboratory protocol study [16].

Figure 2. Firstbeat Bodyguard wearable RR intervals recorded based on ECG signal aquistion with disposable electrodes

D. Signal processing

Firstly, both the data from the PO wrist device and the data from the BG2 reference device were processed with the Firstbeat artifact correction method [17]. Ectopic beats were detected using the algorithm presented by Mateo et al. in [18] and excluded from the evaluation.

Since both devices were not turned on at the same exact moment, the streams of data were synchronized with each other by minimizing their mean absolute difference.

Afterwards, to compensate for eventual time drifts between PO and BG2 clocks, we split the data in intervals of five minutes and performed a new synchronization for each interval.

Using the synchronized PO and BG2 data from each interval, we determined the percentage of correctly detected beats (true positive), extra beats (false positive), and missed beats (false negative). For every PO detected beat, we check how many reference beats were detected in the interval $[t - 0.5l, t + 0.5l]$, where t is the time when the beat was detected and l is the length of the corresponding RR interval. If there is only one reference beat within the interval, then it is considered detected correctly. If there are more than one reference beats, then PO was considered to have missed a beat detection. And if there is no corresponding reference beat, then PO detected a wrong beat. We present an example of this method in Figure 3. For the beat at position $t = 3050$ ms, there are two corresponding reference beats, so we assume that in this case we miss a beat. For the beat at position $t = 5500$ ms, there is no corresponding reference beat, so we consider this an extra beat. This is not a 100% accurate method for beat identification, but, to our knowledge, there is no other better automatic way of doing this.

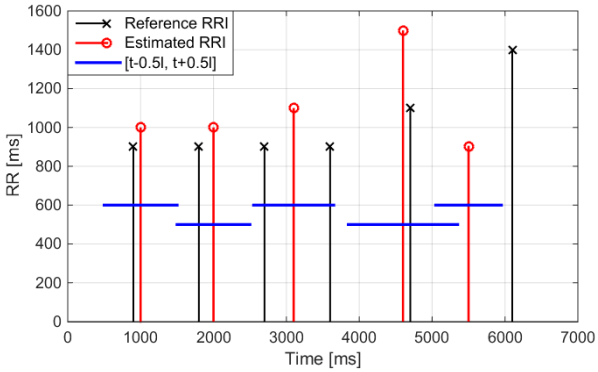


Figure 3. Illustartive example of detecting extra and missing beats

Besides the extra detected and missed beats, we determined the mean absolute error (MAE), the mean absolute percentage error (MAPE), and the root mean square of successive differences (RMSSD).

Finally, we put together the results from all five-minute intervals to obtain the statistics for the whole measurement. Because the five-minute intervals which contain extra or missed beats also contain more artifacts, in order to reduce

the effect of outliers on the analysis, we did not consider them when computing the MAE, MAPE, and RMSSD.

III. Results

The size of the used dataset was 223524 heart beats. The statistics for both uncorrected and corrected data are presented in Table I.

TABLE I. USED DATASET

Error type	Dataset statistics	
	Before artifact correction	***After artifact correction***
Heart beats	223524	221390
Mean [ms]	1071	1062
Std [ms]	244.16	257.93
Min. value [ms]	270	399
Max. value [ms]	2977	2335

Table II provides the summary of the beat detection accuracy, for the cases before and after artifact correction.

TABLE II. BEAT DETECTION ANALYSIS

Error type	Beat–to–beat detection	
	Before artifact correction	***After artifact correction***
Correct beats [%]	99.42	99.57
Extra beats [%]	1.93	0.72
Missing beats [%]	0.58	0.43

The results show that PO detects correctly 99.42% of the heart beats, relative to the BG2 reference, but also adds some extra beats due to movement artefacts. After artifact correction, the amount of false positive beats is reduced from 1.93% to 0.72% (a relative decrease of 62.7%) and the amount of false negative beats is reduced from 0.58% to 0.43% (a relative decrease of 28.3%). This leads to a final detection rate of 99.57%.

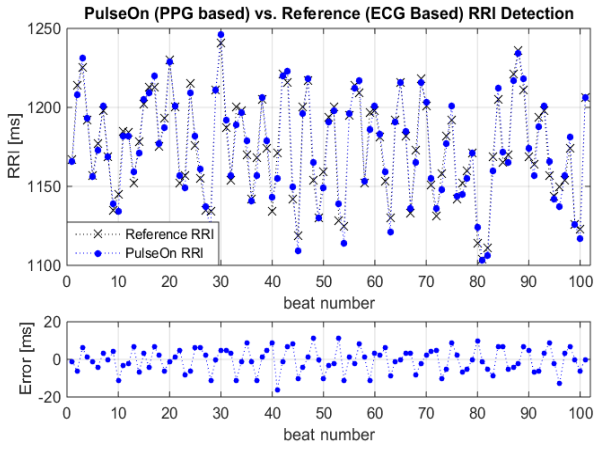


Figure 4. Example showing the RR intervals for 100 heart beats, for POn and BG2. The lower graphs shows the instantenuous error between synchronous intervals

B. *Beat–to–beat interval measurement*

After identifying the correctly detected heart beats, we used them to analyze the accuracy of the beat-to-beat interval estimation. In Figure 4, we show 100 consecutive heart beats, estimated using the PO device and BG2 reference. The top part of the figure shows the duration of the RR-intervals, and the bottom part shows the difference between synchronous PPG-ECG pairs.

The comparison between synchronous beat-to-beat intervals (for all PO detected beats that have only one corresponding reference beat) is performed by evaluating the mean absolute error and the mean absolute percentage error, and the results are given in Table III.

TABLE III. INTERVAL DETECTION STATISTICS

Error type	Beat–to–beat interval estimation	
	Before artifact correction	***After artifact correction***
ME [ms]	-0.32	-0.33
Error std [ms]	14.40	11.74
MAE [ms]	6.68	5.94
MPE [%]	-0.03	-0.03
MAPE [%]	0.62	0.56

The overall mean error is -0.32±14.40 ms before artifact correction and -0.33±11.74 ms after artifact correction. This information is also presented in the Bland-Altman plot from Figure 5. In addition, this figure shows the error distribution and the distribution of the RR interval duration. (Because of the high similarity of the Bland Altman plots for uncorrected and corrected data, we only show the figure for the corrected data.)

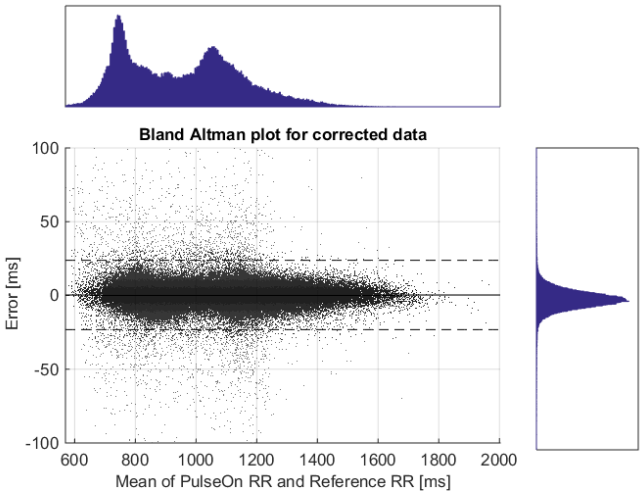


Figure 5. Bland - Altman plot comparing the reference ECG-obtained RR intervals to the PPG-obtained RR intervals, for artifact corrected data. The confidence interval ($\mu \pm 2\sigma$, depicted by the dashed lines) is $[-23.15, 23.83]$ ms.

C. Heart rate variability parameters comparison

The mean error and the mean percentage error provide us with information about how accurately we determine the duration of inter-beat intervals. However, by themselves, they do not provide information about the heart rate variability or about biased measurements.

One measure which describes HRV is the root mean square of successive differences (RMSSD) [4]. We computed it for both PO and BG2 measurements[Table IV]. The RMSSD difference between the PO measurements and the BG2 reference was 4.2 ms (7.00%) for the uncorrected data, and 3.1 ms (4.74%) for the corrected data.

TABLE IV. RMSSD STATISTICS

PulseOn (PPG) RMSSD [ms]	64.18	68.48
Reference (ECG) RMSSD [ms]	59.98	65.38
RMSSD difference [ms]	4.2	3.1

In Table V, we provide several examples of parameters deduced using the RR intervals from this study. The values were determined using the Firstbeat Sports software and represent the average for all the subjects. As the recordings were done during the night, the relaxation time is considerably higher than the stress time, the heart rate is low, the training effect is minimal, and the recovery index is high.

TABLE V. HEART RATE VARIABILITY PARAMETERS

	PulseOn	***Reference***
Relaxation time (min)	195.38	196.31
Stress time (min)	74.53	82.53
Average HR (bpm)	55.84	55.61
Training effect (1->5)	1.03	1.02
Scaled Firstbeat Recovery index (%)	100	100

IV. CONCLUSION

This study explores the accuracy of RR interval detection using PPG based wrist worn device, PO. PO correctly detected 99.57% of the heart beats, and had 0.72% extra beat detections due to movement artefacts during sleep. The MAE was 5.94 ms, and the RMSSD difference against ECG based BG2 reference 3.1 ms. As expected, correcting the artifacts with the Firstbeat Sports software based artefact correction algorithm led to more precise estimation of the beat-to-beat intervals, a very noticeable improvement being visible in the reduction of extra-detected beats.

The results demonstrate that new PPG based HR monitors are becoming a real option for consumer use, not only for HR monitoring while exercising, but also for HRV analysis. PPG provides a more comfortable solution, as they do not require electrodes to be placed on the body. PO device evaluated in this study provides HRV accuracy comparable to ECG based devices and is sufficient for reliable HRV monitoring during sleep.

REFERENCES

[1] J. Sztajzel, "Heart rate variability: a noninvasive electrocardiographic method to measure the autonomic nervous system," *Swiss Medical Weekly*, vol. 134, pp. 514-522, 2004.

[2] T. Myllymäki, H. Rusko, H. Syväoja, T. Juuti, M.-L. Kinnunen, and H. Kyröläinen, "Effects of exercise intensity and duration on nocturnal heart rate variability and sleep quality," *European Journal of Applied Physiology*, vol. 112, pp. 801-809, 2012.

[3] Firstbeat Technologies Ltd., "Stress and Recovery Analysis Method Based on 24-hour Heart Rate Variability", (whitepaper), 2014.

[4] M. Malik, "Heart rate variability," *European Heart Journal Trans. Neural Networks*, vol. 17, pp. 354–381, 1996.

[5] U. Rajendra Acharya, K. Paul Joseph, N. Kannathal, Choo Min Lim, and Jasjit S. Suri, "Heart rate variability: a review." *Medical & Biological Engineering & Computing*, vol. 44, pp. 1031-1051, 2006.

[6] M. Weippert, M. Kumar, S. Kreuzfeld, D. Arndt, A. Rieger, and R. Stoll, "Comparison of three mobile devices for measuring R–R intervals and heart rate variability: Polar S810i, Suunto t6 and an ambulatory ECG system," *European Journal of Applied Physiology*, vol. 109, pp.779-786, 2014.

[7] M. Kingsley, M. J. Lewis, and R. E. Marson, "Comparison of Polar 810 s and an Ambulatory ECG System," International Journal of Sports Medicine, vol. 26, pp. 39 – 43, January 2004.

[8] L. C. Vanderlei, R. A. Silva, C. M .Pastre, F. M. Azevedo, and F. M. Godoy, "Comparison of the Polar S810i monitor and the ECG for the analysis of heart rate variability in the time and frequency domains," *Brazilian Journal of Medical and Biological Research*, vol. 41, pp. 854 – 859, October 2008.

[9] J. Kristiansen, M. Korshøj. J. H. Skotte, T. Jespersen, K. Søgaard, K. Mortensen, and A. Holtermann, "Comparison of two systems for long-term heart rate variability monitoring in free-living conditions - a pilot study," *BioMedical Engineering OnLine*, pp. 10 – 27, April 2011.

[10] Ch. Chuang, J. Ye, W. Lin, K. Lee, and Y. Tai, "Photoplethysmography variability as an alternative approach to obtain heart rate variability information in chronic pain patient," *Journal of Clinical Monitoring and Computing*, (published online), February 2015.

[11] S. Lu, H. Zhao, K. Ju, K. S. Shin, M. H. Lee, K. Shelley, and K. H. Chon, "Can photoplethysmography variability serve as an alternative approach to obtain heart rate variability information?," *Journal of Clinical Monitoring and Computing*, vol. 22, pp.23–29, 2008.

[12] P. Dehkordi, A. Garde, W. Karlen, D. Wensley, J. M. Ansermino, and G.A Dumont, "Pulse rate variability compared with heart rate variability in children with and without sleep disordered breathing," in *Proc. 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Osaka, 2013, pp.6563-6566.

[13] S. Arberet, M. Lemay, P. Renevey, J. Sola, O. Grossenbacher, D. Andries, C. Sartori, and M. Bertschi, "Photoplethysmography-based ambulatory heartbeat monitoring embedded into a dedicated bracelet," in *Proc. Computing in Cardiology Conference (CinC)*, Zaragoza, 2013, pp.935-938.

[14] P. Renevey, J. Sola, P. Theurillat, M. Bertschi, J. Krauss, D. Andries, and C. Sartori, "Validation of a wrist monitor for accurate estimation of RR intervals during sleep," in *Proc. 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Osaka, 2013 pp.5493-5496.

[15] A. Schäfer and J. Vagedes, "How accurate is pulse rate variability as an estimate of heart rate variability? A review on studies comparing photoplethysmographic technology with an electrocardiogram", *International Journal of Cardiology*, vol. 166, pp. 15 – 29, Jun 2013.

[16] J. Parak and I. Korhonen, "Accuracy of Firstbeat Bodyguard 2 beat-to-beat
heart rate monitor," (whitepaper), 2013.

[17] Saalasti S et al. (2004). Artefact Correction for Heartbeat Interval Data. Advanced Methods for Processing Bioelectrical Signals. Available: http://www.firstbeat.com/userData/firstbeat/download/saalasti_et_al_probisi_2004_congress.pdf

[18] J. Mateo and P. Laguna, "Analysis of heart rate variability in the presence of ectopic beats using the heart timing signal," *IEEE Transactions on Biomedical Engineering*, vol. 50, pp.334-343, 2003.

PUBLICATION IV

TOWARDS 24/7 CONTINUOUS HEART RATE MONITORING

by

Tarniceriu, A., Parak, J., Renevey, P., Nurmi, M., Bertschi, M.,
Delgado-Gonzalo, R. & Korhonen, I. (2016)

38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 186-189

Towards 24/7 Continuous Heart Rate Monitoring

Adrian Tarniceriu, Jakub Parak, *IEEE Student Member*, Philippe Renevey, Marko Nurmi, Mattia Bertschi, Ricard Delgado-Gonzalo, and Ilkka Korhonen, *IEEE Senior Member*

***Abstract*— Heart rate (HR) and HR variability (HRV) carry rich information about physical activity, mental and physical load, physiological status, and health of an individual. When combined with activity monitoring and personalized physiological modelling, HR/HRV monitoring may be used for monitoring of complex behaviors and impact of behaviors and external factors on the current physiological status of an individual. Optical HR monitoring (OHR) from wrist provides a comfortable and unobtrusive method for HR/HRV monitoring and is better adhered by users than traditional ECG electrodes or chest straps. However, OHR power consumption is significantly higher than that for ECG based methods due to the measurement principle based on optical illumination of the tissue. We developed an algorithmic approach to reduce power consumption of the OHR in 24/7 HR trending. We use continuous activity monitoring and a fast converging frequency domain algorithm to derive a reliable HR estimate in 7.1s (during outdoor sports, in average) to 10.0s (during daily life). The method allows >80% reduction in power consumption in 24/7 OHR monitoring when average HR monitoring is targeted, without significant reduction in tracking accuracy.**

I. Introduction

Heart rate (HR) and HR variability (HRV) are controlled by the autonomous nervous system and are modified by both internal (e.g. mental stress, relaxation, sleep, alertness) or external (e.g. physical load / activity, posture, etc.) factors. HR/HRV provide rich information about the physical, mental, and health status of an individual. Wearable HR monitoring based on chest straps has been used during physical exercise to facilitate the control of exercise intensity and training effect. Wearable HR/HRV monitoring has also been widely used for objective monitoring of physical activity [1] and stress and recovery [2]. However, the use of a chest-strap or ECG electrodes can cause discomfort, reducing the device's usability and used acceptance, and especially long term adherence. If designed as a wristband, optical HR (OHR) devices do not suffer from this drawback, representing a less obtrusive and more comfortable alternative for HR monitoring. Today, the best OHR devices provide accuracy comparable to ECG based methods for HR during sports [3] and even HRV during low motion interference [4]. Hence, OHR monitoring would offer an attractive method for 24/7 monitoring of behaviors, physical activity, stress, and health. Unfortunately, due to its measurement principle based on optical illumination of the tissue, the inherent power consumption of the OHR technology is significantly higher than that of the ECG based methods.

Our objective was to develop an algorithmic approach to reduce power consumption of the OHR technology in 24/7 HR monitoring. In this approach, HR is sampled semi-continuously, and continuous activity monitoring and fast adapting frequency domain estimation is used for fast convergence of the algorithm to provide a reliable HR estimate during various activities. The method was evaluated during sports, daily life and sleep.

II. Optical Heart Rate Monitoring

Optical HR monitoring is based on the photoplethysmography (PPG) principle. Light emitted by a LED is transmitted at the surface of the body tissue. During the propagation, the light-wave suffers reflection, refraction, scattering, and absorption and the resulting signal is detected by a photodetector (PD) [5]. The PD can detect reflected light (reflectance mode) or back-scattered light (transmission mode). Given that the received light intensity depends on the variations of the subcutaneous blood flow, and as these variations are directly related to heart pulsations, we can use the detected signal to estimate the heart rate.

One of the main problems of PPG measurement is that the useful signal is corrupted by ambient light and other electro-magnetic radiations (ambient light artefacts), by gravity and by voluntary and involuntary subject movements (motion artefacts). The ambient light artefacts influence can be measured using multiplexing techniques and eliminated by subtractive techniques [6]. An efficient way to reduce the motion artefacts is to use a motion reference signal provided by an accelerometer and to perform signal enhancement afterwards [7]. In this way, we may obtain reliable HR estimates even under intense physical activities.

The OHR device used in this study is the PulseOn device (PulseOn, Espoo, Finland). It is a wearable wristband consumer OHR monitor which uses two light wavelengths (green and infrared) and has optimally matched LED-PD distances to allow the measurement of blood flow in the wrist. Both the mechanical casing and the strap are designed to provide a stable sensor-skin contact, reducing the artefacts [3, 4]. The HR detection algorithm applies the accelerometer data to reduce the motion artefacts and provide accurate HR estimation for a range of activities from rest or daily office routine to intensive training. Added to this, the acceleration data is used to determine activity related parameters such as

A. Tarniceriu is with PulseOn SA, Jacquet-Droz 1, 2002, Neuchâtel, Switzerland; e-mail : adrian.tarniceriu@pulseon.com.

J. Parak and I. Korhonen are with the Department of Signal Processing, Tampere University of Technology, Tampere, Finland, and PulseOn Oy, Espoo, Finland.

Marko Nurmi is with PulseOn Oy, Espoo, Finland.

R. Delgado-Gonzalo, Ph Renevey, and M. Bertschi are with CSEM - Centre Suisse d'Electronique et Microtechnique, Jacquet-Droz 1, 2002 Neuchâtel, Switzerland

step or calory count. The device can monitor HR, performed activity, PPG and accelerometer signals.

III. Power Consumption Improvement

Most algorithms for HR detection from PPG are based on frequency domain estimation of the HR frequency in a noisy PPG signal. As the PPG signal-to-noise ratio during motion is usually poor (even below 1:100) and the signal may include several rhythmic components close to the HR frequency due to e.g. physical motion artefact, the convergence of the algorithms is usually slow.

We developed a semi-continuous OHR monitoring algorithm ("sampled HR"). In this approach, acceleration is continuously monitored to estimate the activity status, to predict the expected HR level based on activity, and to continuously estimate the motion artefact frequency. The OHR sensor is sampled semi-continuously in pre-determined intervals (e.g., every 60s or 300s, depending on the application). When OHR is sampled, the initial HR is estimated from the activity to speed up convergence, and the frequency content of the motion is used to filter out activity related frequencies from the HR spectrum. HR is estimated until a reliable HR is achieved or when timeout (*P*) is reached (the HR reliability is estimated based on spectral separation of HR and motion related signals in PPG). In this study, we chose N = 60 seconds, as one HR estimate per minute still provides a reliable description of the heart activity and training effect and maintains HR trend information over 24/7. We defined an indicator for the HR estimation reliability and follow the next steps:

- Set *P* to a value higher than the algorithm convergence time (e.g., 20 seconds).
- Every *N* seconds, repeat:
    1. Start PPG acquisition and optimize PPG acquisition parameters for current ambient light and motion conditions;
    2. Initialize HR algorithm with predicted HR based on activity status;
    3. For each new sample, estimate HR and iterate algorithm to remove motion artefact and adapt frequency domain estimator to HR frequency;
    4. If a reliable HR value is found, return the value and stop estimation;
    5. If no reliable HR value is found after *P* seconds, return the latest HR estimate.

A visual description of the algorithm is given in Figure 1. In the estimation interval starting from 0, a reliable HR is found after 9 seconds. In the interval starting from 60 seconds, there is no reliable HR found. In this case, the algorithm returns the last estimated value, represented by the blue square.

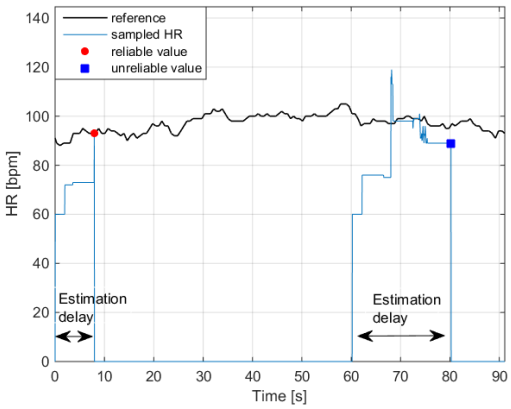


Figure 1. Heart-rate estimation example for the proposed algorithm

IV. Experimental Validation

To validate the algorithm, we compare the performances of the sampled algorithm against continuous OHR estimation and ECG based reference during sports, daily life and sleep. As the sampled-mode algorithm was designed with the aim of faster convergence, we might expect that its performance is not as high as for the continuous-mode algorithm. But this should be compensated by reduced power requirements.

For performance estimation, we compute the following parameters:

- *Mean Absolute Error (MAE)*: average of the absolute difference between the reference and the estimated HR.
- *Reliability*: percentage of time when the absolute error is below 10 beats per minute (bpm).
- *Accuracy*: 100 – mean absolute percentage error.

In addition, for the sampled mode, we compute

- *Reliability rate*: the percentage of 60-second intervals for which a reliable HR value was found.
- *Estimation delay*: the average time duration required to obtain a HR value.
- *Reliable estimation delay*: the average time duration required to obtain a reliable HR value.

These measures will be computed for four different datasets, covering a wide range of activities.

The experimental procedures described in the following comply with the principles of the Helsinki Declaration of 1975, as revised in 2000. All subjects gave informed consent to participate and they had the right to withdraw from the study at any time.

1) Controlled Laboratory Protocol

The test group consists of 19 volunteers, 9 male and 10 female, 28.3±5.69 years old, non-smokers. All subjects perform moderate physical activities weekly. Each subject followed a protocol including rest, walking and running on a treadmill, and ergo-cycling (Table I). The PPG was monitored using the PulseOn device, and the ECG-based Polar Electro RS800CX [8] was used as reference.

TABLE I. LABORATORY PROTOCOL

Activity	Duration
Standing	1min
Walking on a treadmill at 3km/h, 0% inclination	3min
Walking on a treadmill at 3km/h, 5% inclination	3min
Walking on a treadmill at 3km/h, 10% inclination	3min
Walking on a treadmill at 5km/h, 0% inclination	3min
Walking on a treadmill at 5km/h, 5% inclination	3min
Walking on a treadmill at 5km/h, 10% inclination	3min
Running on a treadmill at 9km/h, 0% inclination	3min
Running on a treadmill at 11km/h, 0% inclination	3min
Rest sitting	4min
Cycling 60rpm*	3min
Cycling 90rpm*	3min
Rest sitting	4min

2) Outdoor Activities

The outdoor activities consist of walking, running, and cycling, both on and off-road. The test group contains 28 recordings made by 9 subjects, 8 male and 1 female, 33.5±10.3 years old, with a total duration of 21.3 hours. The used reference was either Polar Electro RS800CX or Firstbeat Bodyguard 2, both ECG-based. PPG was recorded with the PulseOn device.

3) Sleep

The test group consists of 10 volunteers, 8 male and 2 female, 35.9±10.3 years old, non-smokers. All subjects perform moderate physical activities weekly. A total of 13 recordings were made. Firstbeat Bodyguard 2 was used as reference and PPG was recorded with the PulseOn device. Subjects performed the recordings at their homes in normal bedroom sleeping conditions. The average non-stop recorded sleep time of all subjects was 5.1±1.2 hours, and the total recording duration is 65.2 hours.

4) Daily Activities

These activities consist of daily office or house work. The used reference was either Polar Electro RS800CX or Firstbeat Bodyguard 2. The PPG signal was recorded with the PulseOn device. These recordings were made by three subjects and have a total duration of 17 hours.

All the analysis were performed offline. HR was derived from PPG with PulseOn's PPG algorithm and with the sampled mode algorithm as presented above.

V. RESULTS

The performance metrics for each dataset are summarized in Table II. The sampled mode algorithm resulted in slightly higher MAE and lower reliability during sports but lower MAE and higher reliability during daily activity and sleep. The average reliable estimation delay varied from 7.1s during outdoor sports to 10.0s during daily life. Figures 2-5 show one example of continuous and sampled-mode HR estimation for each dataset. The black line represents the reference and the green line is the continuous mode estimate. The red dots are sampled mode estimates considered reliable. The blue dots indicate cases when no reliable HR was found after 20 seconds (even so, the estimated values are still close to the reference most of the time).

VI. CONCLUSIONS

We developed and evaluated an algorithm for semi-continuous OHR monitoring during various activities. The results show that very low power semi-continuous OHR trending is possible without sacrificing the accuracy of HR detection as compared to continuous monitoring. For sports, the sampled-mode performance is below the continuous-mode performance, but this was expected: continuous HR tracking is able to correct some errors caused by the motion, while in sampled mode this is not possible. The accuracy difference is below 1%, and the MAE difference is below 1 bpm (note that even the sampled-mode performance is better than other optical heart rate monitors [4, 9]). The average

TABLE II. PERFORMANCE METRICS FOR THE LAB PROTOCOL, OUTDOOR ACTIVITIES, SLEEP, AND DAILY ACTIVITIES

	Lab protocol		Outdoor		Sleep		Daily	
N subjects	19		9		10		3	
Duration [hours]	12.3		21.3		65.2		17	
	Continuous mode	Sampled mode	Continuous mode	Sampled mode	Continuous mode	Sampled mode	Continuous mode	Sampled mode
MAE [bpm]	3.4	4.2	3.1	4.5	1.8	1.3	3.3	2.9
Reliability [%]	93.9	92.4	92.8	89.1	98.5	99.5	93.2	94.7
Accuracy [%]	97.1	96.3	97.4	96.6	96.6	97.6	95.4	95.9
Estimation delay[s]	-	8.6	-	8.1	-	9.4	-	12.2
Reliable estimation delay[s]	-	7.8	-	7.1	-	9.3	-	10.0
Reliability rate [%]	-	89.6	-	89.8	-	94.4	-	77.1

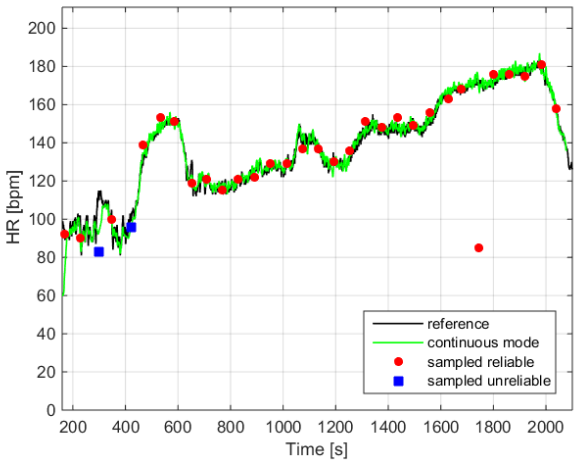


Figure 2. Heart-rate estimation example for the laboratory protocol

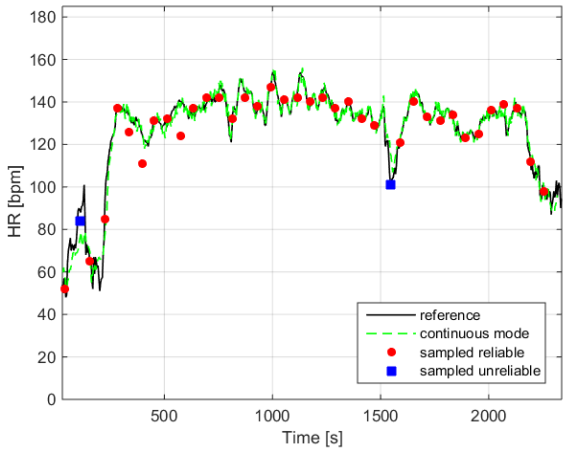


Figure 3. Heart-rate estimation example for outdoor activities (walk, run, and short break)

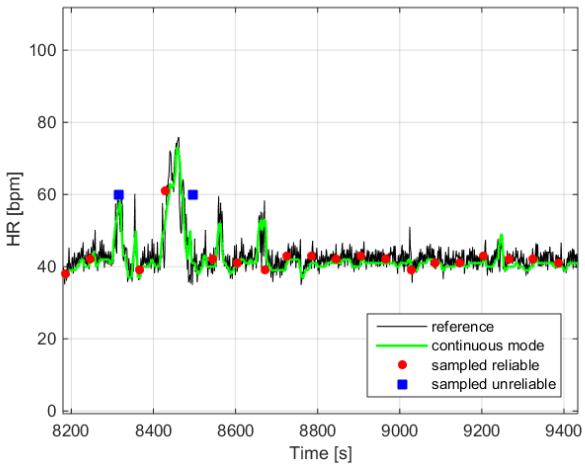


Figure 4. Heart-rate estimation example for sleep

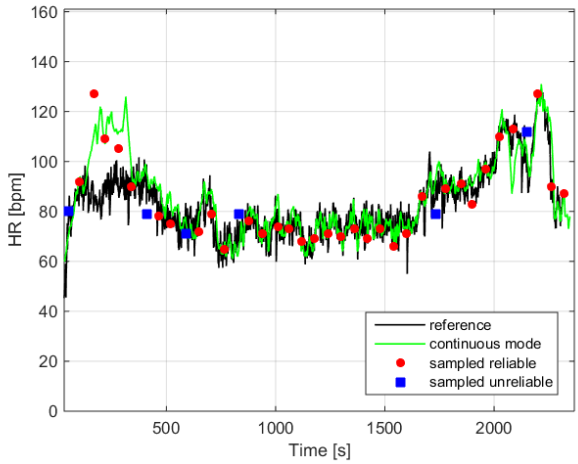


Figure 5. Heart-rate estimation example for daily activities

estimation duration is 8.6 seconds for the laboratory protocol, 8.1 seconds for outdoor activities, 9.4 seconds for the sleep recordings, and 12.2 seconds for daily activities. If we estimate the heart rate every minute, this means a reduction of 85.7%, 86.5%, 84.3%, and 79.7% of the optical chain power consumption, respectively; for longer sampling intervals, the savings are even more significant. Future work will focus on further reducing these durations, but the current results already allow extending OHR monitoring towards real 24/7 use without sacrificing accuracy.

REFERENCES

[1] S. Mutikainen, E. Helander, J. Pietilä, I. Korhonen, U.M. Kujala, "Objectively measured physical activity in Finnish employees: a cross-sectional study, " *BMJ Open*, 4:e005927. doi:10.1136/bmjopen-2014-005927, 2014.

[2] J. Pietilä, E. Helander, T Myllymäki, I. Korhonen, H Jimison, M Pavel, "Exploratory analysis of associations between individual lifestyles and heart rate variability –based recovery during sleep," in *37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Milan, 2015

[3] R. Delgado-Gonzalo, J. Parak, A. Tarniceriu, P. Renevey, M. Bertschi, and I. Korhonen, "Evaluation of Accuracy and Reliability of PulseOn Optical Heart Rate Monitoring Device," in *Proc. 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Milan, 2015, pp. 430-433.

[4] J. Parak, A. Tarniceriu, P. Renevey, M. Bertschi, R. Delgado-Gonzalo, and I. Korhonen, "Evaluation of Beat-to-Beat Detection Accuracy of PulseOn Wearable Optical Heart Rate Monitor," in *Proc. 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Milan, 2015, pp. 8099-8102.

[5] M. Bertschi, P. Renevey, J. Solà, M. Lemay, J. Parak and I. Korhonen, "Application of optical heart rate monitoring," in *Wearable Sensors. Fundamentals, Implementation and Applications*, Elsevier Academic Press, 2014, p. 656.

[6] L. Rossini, R. Vetter, C. Verjus, P. Theurillat, P. Renevey, M. Bertschi, J. Krauss, "Robust ear located heart rate monitor," in *Proc. 2nd International Conference on Biomedical Electronics and Devices*, Porto, 2009, pp. 214-219..

[7] S. Vaseghi, *Advanced Digital Signal Processing and Noise Reduction, 4th Edition*. Wiley, 2009.

[8] M. Kingsley, M. Lewis and R. Marson, "Comparison of Polar 810s and an ambulatory ECG system," *International Journal of Sports Medicine*, vol. 26, pp. 39-43, 2005

[9] J. Parak and I. Korhonen, "Evaluation of wearable consumer heart rate monitors based on photopletysmography," in *Proc. 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Chicago, USA, 2014, pp. 3670-3673.

PUBLICATION V

ESTIMATING HEART RATE, ENERGY EXPENDITURE, AND PHYSICAL PERFORMANCE WITH A WRIST PHOTOPLETHYSMOGRAPHIC DEVICE DURING RUNNING

by

Parak, J., Uuskoski, M., Machek, J. & Korhonen, I. (2017)

JMIR mHealth and uHealth, 5(7), e97

Original Paper

Estimating Heart Rate, Energy Expenditure, and Physical Performance With a Wrist Photoplethysmographic Device During Running

Jakub Parak[1,2], MSc; Maria Uuskoski[2,3], MSc; Jan Machek[2], MSc; Ilkka Korhonen[1,2], PhD

[1]BioMediTech Institute, Faculty of Biomedical Sciences and Engineering, Tampere University of Technology, Tampere, Finland

[2]PulseOn Oy, Espoo, Finland

[3]Department of Biology of Physical Activity, University of Jyväskylä, Jyväskylä, Finland

Corresponding Author:
Jakub Parak, MSc
BioMediTech Institute
Faculty of Biomedical Sciences and Engineering
Tampere University of Technology
Korkeakoulunkatu 10
Tampere, 33720
Finland
Phone: 358 447647457
Email: jakub.parak@tut.fi

Abstract

Background: Wearable sensors enable long-term monitoring of health and wellbeing indicators. An objective evaluation of sensors' accuracy is important, especially for their use in health care.

Objective: The aim of this study was to use a wrist-worn optical heart rate (OHR) device to estimate heart rate (HR), energy expenditure (EE), and maximal oxygen intake capacity (VO_{2Max}) during running and to evaluate the accuracy of the estimated parameters (HR, EE, and VO_{2Max}) against golden reference methods.

Methods: A total of 24 healthy volunteers, of whom 11 were female, with a mean age of 36.2 years (SD 8.2 years) participated in a submaximal self-paced outdoor running test and maximal voluntary exercise test in a sports laboratory. OHR was monitored with a PulseOn wrist-worn photoplethysmographic device and the running speed with a phone GPS sensor. A physiological model based on HR, running speed, and personal characteristics (age, gender, weight, and height) was used to estimate EE during the maximal voluntary exercise test and VO_{2Max} during the submaximal outdoor running test. ECG-based HR and respiratory gas analysis based estimates were used as golden references.

Results: OHR was able to measure HR during running with a 1.9% mean absolute percentage error (MAPE). VO_{2Max} estimated during the submaximal outdoor running test was closely similar to the sports laboratory estimate (MAPE 5.2%). The energy expenditure estimate (n=23) was quite accurate when HR was above the aerobic threshold (MAPE 6.7%), but MAPE increased to 16.5% during a lighter intensity of exercise.

Conclusions: The results suggest that wrist-worn OHR may accurately estimate HR during running up to maximal HR. When combined with physiological modeling, wrist-worn OHR may be used for an estimation of EE, especially during higher intensity running, and $VO_{2Max,}$ even during submaximal self-paced outdoor recreational running.

(JMIR Mhealth Uhealth 2017;5(7):e97) doi:10.2196/mhealth.7437

KEYWORDS

fitness trackers; photoplethysmography; heart rate; heart rate determination; exercise test; oxygen consumption; energy metabolism

Introduction

Advances in wearable sensors enable long-term monitoring of health and wellbeing indicators in various conditions and activities in both consumers and patients. Recently, significant progress in the size, power consumption, and accuracy of various different sensing technologies has led to an introduction of affordable wearable sensors with a reasonable battery life and

capability to monitor, for example, physical activity, sleep, heart function, and so on. However, the reliability and accuracy of the produced information has been questioned and significant differences between different brands have been found [1]. Therefore, an objective scientific evaluation of available wearable sensors is essential for the progress of their use, especially for health applications such as chronic disease prevention and management.

Heart rate (HR) monitoring provides valuable information on physiology and health status during sports, daily life, and sleep. Chest strap HR monitors have been used during sports to quantify and control training loads since the late 1980s. The main limitation for the wide and long-term use of chest strap HR monitors, especially in female users, is the discomfort that is caused by the tightness of the chest strap and possible skin irritations. Therefore, their application has remained relatively limited, especially in real-life wearable monitoring.

Wearable optical HR (OHR) monitoring technology based on photoplethysmography (PPG) has been significantly improved recently because of miniaturized low-power hardware and improved embedded algorithms. OHR technology can be applied on almost any part of the body, such as on the wrist, and can hence overcome some challenges of chest strap HR monitors in their usability and long-term use. However, relatively few scientific studies have reported OHR technology performance and accuracy in laboratory or real-life conditions. Olenick et al evaluated a Mio Alpha wrist OHR device during a graded treadmill exercise test until volitional fatigue and found a strong correlation between OHR and ECG-based HR [2]. In a study by Parak and Korhonen [3], wrist and forearm OHR devices were evaluated during multiple physical activities (walking, running, and biking) with a 5% agreement ranging from 76% to 78%. Delgado-Gonzalo et al evaluated the accuracy and reliability of two different wrist OHR devices (PulseOn and Mio Alpha) against ECG-derived HR in laboratory conditions during a wide range of physical activities and found the mean absolute error of PulseOn to be 3% and Mio Alpha to be 6% during laboratory protocol [4]. Similar or better accuracy was seen during normal outdoor sports activities [4]. In general, wrist-worn OHR devices seem to provide good accuracy during running, but less so in some other activities, such as biking and weight lifting [5-7]. These studies suggest that the currently available high-end OHR devices are reaching acceptable accuracy for HR monitoring during cardiovascular sports such as running, while different brands and devices may experience significant differences in their performance.

Exercise HR is itself a valuable parameter. For example, it allows a real time control of training loads. However, exercise HR alone is challenging to interpret for users, and an estimation of more advanced physiological parameters during exercise would be beneficial to allow a more insightful analysis of the training. An estimation of momentary oxygen consumption and total energy expenditure (EE) for each training session and an estimation of changes in physical performance achieved by regular training are examples of these insightful parameters. An indirect calorimeter is one of the most accurate reference methods for estimating EE. This method is based on the analysis of respiratory gases and is commonly used in laboratory settings. HR has also been used for estimating oxygen consumption. Montgomery et al [8] evaluated the accuracy of oxygen consumption and EE estimation based on chest strap HR monitors and found a slight underestimation with a 6% coefficient of a variation of 6% for oxygen consumption and 13% for EE. Keytel et al [9] reported a correlation coefficient of .913 between the chest strap HR-based method and indirect calorimeter-based EE. Running speed can also be used to estimate oxygen consumption; in runners, a strong correlation (>.99) has been reported [10,11]. Robertson et al [12] found a significant correlation between EE estimates based on indirect calorimetry and a HR chest strap based method during low intensity exercise and maximum intensity exercise. However, Wallen et al observed poor accuracy in an EE estimation of four OHR smart watches as compared with indirect calorimetry [13].

Physical performance may be estimated by the maximal oxygen consumption (VO_{2Max}) of a person. VO_{2Max} can be measured directly with an expiratory gas analyzer during a maximal voluntary exercise test. Running speed may also be used to estimate VO_{2Max} [14]. By estimating the oxygen consumption and speed during submaximal exercise, it is possible to estimate VO_{2Max} without maximal exercise testing [15]. LeBoeuf et al found good accuracy of an OHR sensor placed in the ear in the assessment of EE and VO_{2Max}: −0.7 (SD 7.4%) and −3.2 (SD 7.3%) [16]. However, to our knowledge, the accuracy of a wrist-worn OHR on the estimation of EE, oxygen consumption, or VO_{2Max} has not been widely studied.

The objectives of the current study were to use OHR to estimate HR, EE, and VO_{2Max} during running and to evaluate the accuracy of the estimated parameters (HR, EE, and VO_{2Max}) against a chest strap HR and respiratory gas analysis derived from golden reference values.

Methods

Subjects

Twenty-four healthy adults (13 males and 11 females) participated in the study (Table 1). The inclusion criteria were age (18-55 years), BMI (18-30), normal self-reported health status, experience in treadmill running, and a self-estimated ability and willingness to continue the exercise protocol with an increasing load until exhaustion. The health status of the subjects was evaluated in advance through a self-reporting questionnaire and a verbal interview by a trained sports laboratory physiologist about the subjects' capabilities to reach maximum performance. The subjects provided signed informed consent to participate in the study and they were told that they could withdraw from the study or protocol at any time, if they so desired. The study followed the ethical guidelines of the Helsinki declaration.

Table 1. Demographics of the participants.

Parameter	All	Male	Female
No. of participants	24	13	11
Age in years, mean (SD)	36.2 (8.2)	36.8 (9.1)	35.4 (7.2)
Height in cm, mean (SD)	174.1 (8.0)	180.0 (5.6)	167.2 (3.5)
Weight in kg, mean (SD)	69.2 (10.6)	76.1 (9.0)	61.1 (5.2)
BMI[a] in kg/m^2, mean (SD)	22.7 (1.9)	23.4 (1.8)	21.8 (1.7)

[a]BMI: body mass index.

Study Protocol

The study protocol included two parts: (1) a submaximal outdoor running test and (2) a maximal voluntary exercise test in the sports laboratory. The submaximal outdoor running test was performed in regular outdoor conditions in Finland with the aim of providing data from uncontrolled and sometimes challenging conditions, where subjects would train and perform their fitness tests when provided with self-testing equipment, such as a PPG wrist device and a mobile phone. The data from the submaximal outdoor running tests was used to estimate $VO_{2Max,}$ based on wrist PPG and mobile phone GPS data. The maximal voluntary exercise was performed to provide a standardized reference ("ground truth") for VO_{2Max} for each individual and to compare EE from a wrist PPG against a standard respiratory gas analysis-based EE reference during running. The order of the tests was randomized with a maximal time difference of 7 days.

The submaximal outdoor running test was performed on a pre-defined outdoor track with a flat surface. The subjects were instructed to run at a self-determined pace for at least 20 min, targeting moderate to vigorous subjectively assessed intensity, and to run 5 km. HR was monitored with an optical wrist worn heart rate monitor (PulseOn, Espoo, Finland) and GPS data with a mobile phone (Samsung S3 Galaxy Trend). A Polar V800 HR monitor (Polar Electro, Kempele, Finland) with a built-in GPS sensor was used as a reference for the distance. The GPS reference for the distance was necessary, as the subjects performed the actual running test without continuous supervision and, hence, had a possibility to vary their running route to some extent. The PulseOn mobile app was used to track and store HR and running speed during the test. Field tests were performed outdoors between November 2014 and January 2015 in Finland in regular winter training conditions, that is, during days when it was not raining or snowing, the testing track was not too slippery to cause health risks, and the temperature was above -10 °C. The subjects were instructed to wear their own outdoor sports clothing as appropriate for the current weather during the test. These conditions are typical outdoor training conditions in Finland and, hence, provide a good benchmark for challenging real outdoor training conditions that are faced by ordinary citizens while training.

The maximal voluntary exercise test was performed in a sports testing laboratory with a treadmill (OJK-2, Telineyhtymä, Kotka, Finland). The indoor temperature during the tests was 20 °C. During the test, the subjects wore a face mask from the respiratory gas analyzer (Metalyzer 3B, Metasoft Studio 4.8, Cortex Biophysik GmbH, Leipzig, Germany), the PulseOn wrist HR device, and a chest strap HR device (RS800CX, Polar Electro, Kempele, Finland). The treadmill inclination was set to 0.6°. After setting up the measurement devices and instructing the user about the study protocol and the use of the treadmill, the subject performed a warm-up run at 8 km/h for 6 min. Then, the subject stood still for 6 min and the first blood sample was taken, after which the actual test started. The running speed was increased by 1 km/h, which was maintained for 3 min to reach a stable metabolism at each load. The initial running speed was set so that the predicted number of loads that the subject would be able to complete would be between 8 and 10. Between transitions, the treadmill was stopped for 20-30 s, during which a blood sample was drawn from the subject's finger to estimate the blood lactate (Biosen C_Line, EKF Diagnostic, 42 Barleben/Magdeburg, Germany). The test was continued until the subject wanted to stop (a stop signal was agreed upon in advance) or the following end criteria, based on recommendations by the Finnish Sports Testing Society, were met: (1) predicted maximum heart rate was reached, (2) measured VO_2 was stabilized or started to decrease, (3) blood lactate level increased above a threshold, or (4) respiratory exchange ratio was >1.1. After the test, the subject was allowed to recover for 3 min, which was followed by a 7 min cool down jog at a self-selected speed. After this, the final blood sample was taken.

Energy Expenditure and Maximal Oxygen Intake Capacity Estimation From Optical Heart Rate

PulseOn OHRs recorded during submaximal and maximal tests were re-analyzed offline because of the randomized order of the field and laboratory tests. VO_{2Max} was calculated from the submaximal test and EE was calculated from the maximal exercise test. HR, GPS data, and personal subject information (height, weight, gender, and age) were used for calculations. Both maximal HR estimated during the maximal exercise test and maximal HR estimated from the subject's age (208 − 0.7 × age [17]) were used for the VO_{2Max} calculation. VO_{2Max} estimated offline from the submaximal test was used for the EE estimation during the maximal exercise test.

The estimation of total EE was based on a method developed earlier [18]. Neural networks were used to derive momentary oxygen consumption (VO_2) from HR. Differences in the HR-VO_2 relationship during the different exercise phases (on and off phases) were included in the model. Personal maximal HR and estimated VO_{2Max} were used for the calculation of the

momentary VO_2 value. EE was then estimated from VO_2, respiratory quotient (RQ), and caloric equivalent [18]. RQ describes the ratio between carbon dioxide produced and oxygen consumed in metabolism, varying from 0.70 to 1.00. RQ has a well-established deterministic relationship with the caloric equivalent, which describes the amount of energy expended per one liter of consumed oxygen, varying from 4.69 to 5.05 kcal/l O_2 [19]. Both exercise intensity and duration affect the RQ and caloric equivalent. An increase in exercise intensity results in an increased RQ and caloric equivalent, due to the increased oxidation of carbohydrate and decreased oxidation of fat. A prolonged exercise duration has an opposite effect, due to the increased oxidation of fat and decreased oxidation of carbohydrate. When the momentary VO_2 and caloric equivalent are known, it is possible to calculate the momentary EE. The total EE can be calculated by summing up the momentary EE values.

VO_{2Max} was estimated from OHR and GPS speed recorded during the self-paced running test by a company (Firstbeat, Jyväskylä, Finland) [20]. The method is based on a linear relationship between VO_2 and the running speed. First, speed and OHR data are segmented to different HR ranges and the reliability of different data segments is estimated by calculating the correlation between HR and speed and comparing that to the variance of the data in that segment. In case of a wide variance and low correlation, the segment is discarded as being unreliable. Then, the most reliable data segments are used to estimate VO_{2Max} by utilizing the relationship between HR and speed. Finally, VO_{2Max} is estimated as the reliability weighted average of the segments.

Data Analysis

A maximal voluntary exercise test was used to determine the reference ("ground truth") $VO_{2Max,}$ as well as measure EE during the test. EE was measured by averaging the measured EE, based on a respiratory gas analysis for each minute. Equations defined by Weir [21] were used to calculate EE, based on respiratory gas measurements. VO_{2Max} was determined by using criteria defined by the Finnish Society of Sport Sciences [22].

HR data from a chest belt acquired during the laboratory test was analyzed with Firstbeat Sports software (Firstbeat, Jyväskylä, Finland, version 4.5). After applying an artifact correction algorithm to the signals, the maximum HR value was observed. A second-by-second chest strap HR was used as a reference for the OHR signal during the maximal voluntary test, and the acquired maximum HR value was used as the measured maximum HR in the further analysis.

Statistical Analyses

The HR estimation accuracy of the wrist PPG device was estimated during the maximum exercise test by comparing HR from the wrist PPG device with chest strap-based HR. First, the data were re-sampled at 1.5 s sampling intervals. HR signals were synchronized in time by maximizing the cross-correlation between the signals at t=0. Then, the HR data was averaged over 5 s non-overlapping windows. HR accuracy was estimated by the following parameters [3,4].

Reliability: The percentage of time that the absolute error is smaller than 10 bpm.

Accuracy: The complement of the relative error (ie, 100% mean absolute percentage error).

The difference between VO_{2Max} estimated with a wrist PPG device and GPS data during a submaximal test and with a gas analyzer during a maximal exercise test was compared by calculating the bias, mean absolute error (MAE), mean absolute percentage error (MAPE), and correlation coefficient (either Pearson when data was normally distributed or Spearman when this was not the case) between the estimates. Bland-Altman plots were constructed to allow a visual presentation of the agreement between the two estimation methods and their average error (bias), as well as 95% confidence limits of agreement.

The difference between EE estimated from the wrist PPG device and respiratory gas analysis was calculated during the maximum exercise test. The analysis was carried out separately for light intensity (below aerobic threshold) and medium heavy intensity (between aerobic and anaerobic thresholds). The estimation was only performed from light to medium heavy intensity levels, as higher intensity levels can change the body acid-base balance, which can distort the indirect calorimetry method [23]. The aerobic and anaerobic thresholds of the subjects were determined by the guidelines of the Finnish Society of Sport Sciences [22,24]. Bland-Altman plots were generated for a visual analysis of the error, and bias, MAE, MAPE, and correlation coefficients were calculated for the data.

The normal distribution of data was examined by the Shapiro-Wilk test. The difference between the methods was tested with a paired t test in case normal distribution was confirmed and with the Wilcoxon signed rank test when normal distribution could not be confirmed. Pearson correlation coefficient was computed between normally distributed parameters, while Spearman rank correlation coefficient was used for the other parameters not meeting the normal distribution assumption. The strength of the correlation coefficients was interpreted based on the following definitions: weak ($r \leq .5$), moderate (r=.5–.7) and strong ($r \geq .7$). All statistical tests were performed as two-sided and the level of significance was set at $P<.05$.

All data analysis was carried out with MathWorks Matlab (version 8.5). All statistical testing was carried out with IBM SPSS statistics (version 22).

Results

Heart Rate Accuracy During Treadmill Running

HR estimated with a wrist PPG device appeared to closely follow HR monitored with a chest strap (Table 2). In most cases, wrist PPG HR estimated HR accurately over the entire protocol, even up to maximum HR and running speeds, as shown in Figure 1 (parts A and B). In a few cases, there were occasional outliers, as shown in Figure 1 (part C: in the worst case, OHR artifacts during the beginning of the recording are likely related to poor perfusion before fully warming up, while at the end, the subject was struggling to maintain the running speed, resulting in

non-rhythmic hand motions because the subject was aiming to gain support from the treadmill handles.). This can also be seen in Figure 2, which presents the Bland-Altman plot of the HR during the entire laboratory protocol from a wrist OHR device and chest strap HR.

Figure 1. Comparison of HR from chest strap (black line) and wrist PPG device (red line) during maximum exercise test: (A) best accuracy, (B) average accuracy, and (C) worst case.

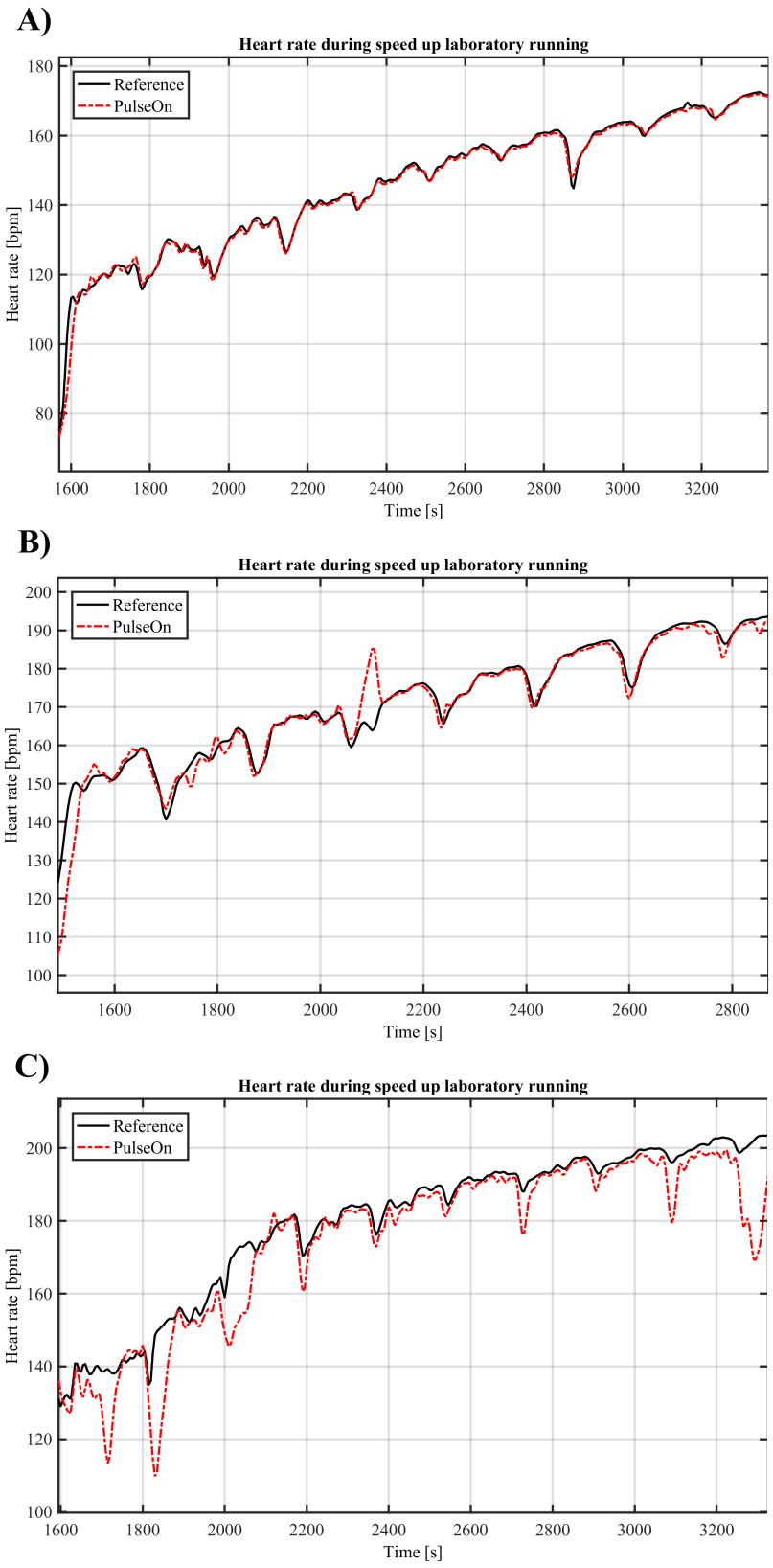


Figure 2. Bland−Altman plot comparing the wrist PPG device and chest strap HR device during maximum exercise protocol in all 24 subjects (solid horizontal line: bias, dashed lines: 95% confidence limits of agreement).

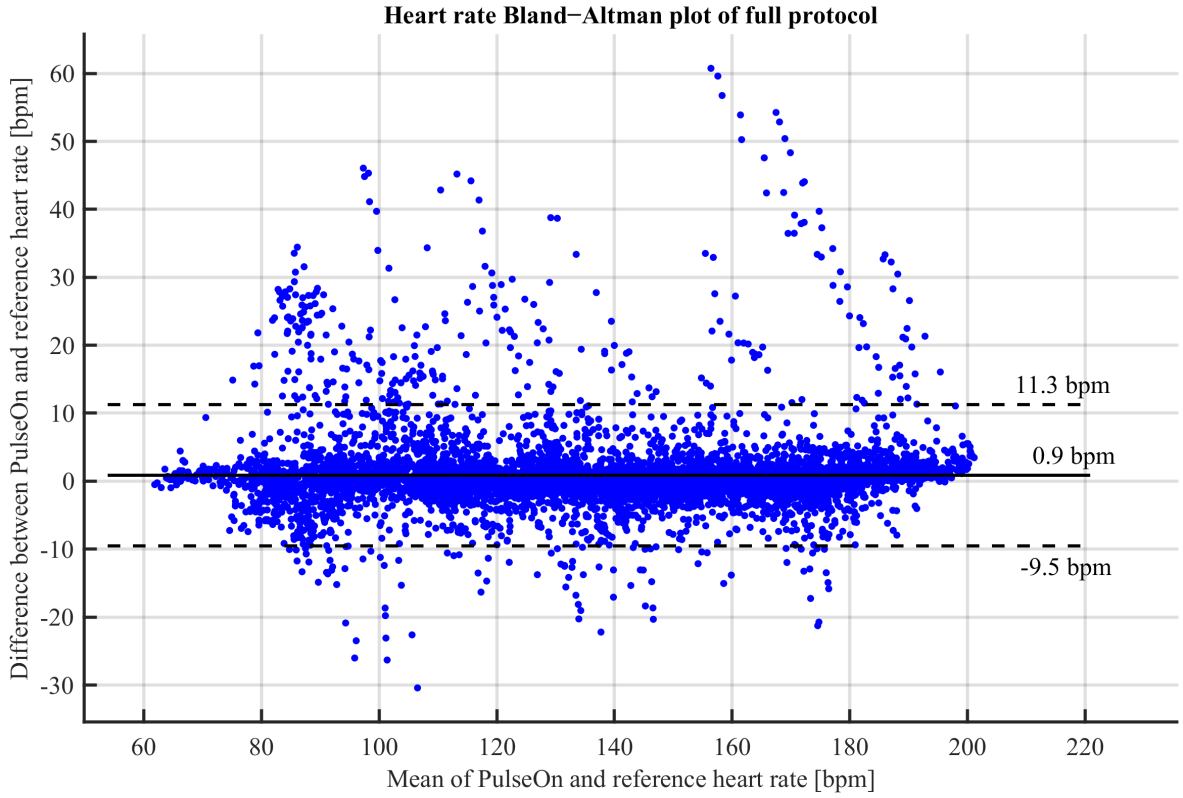
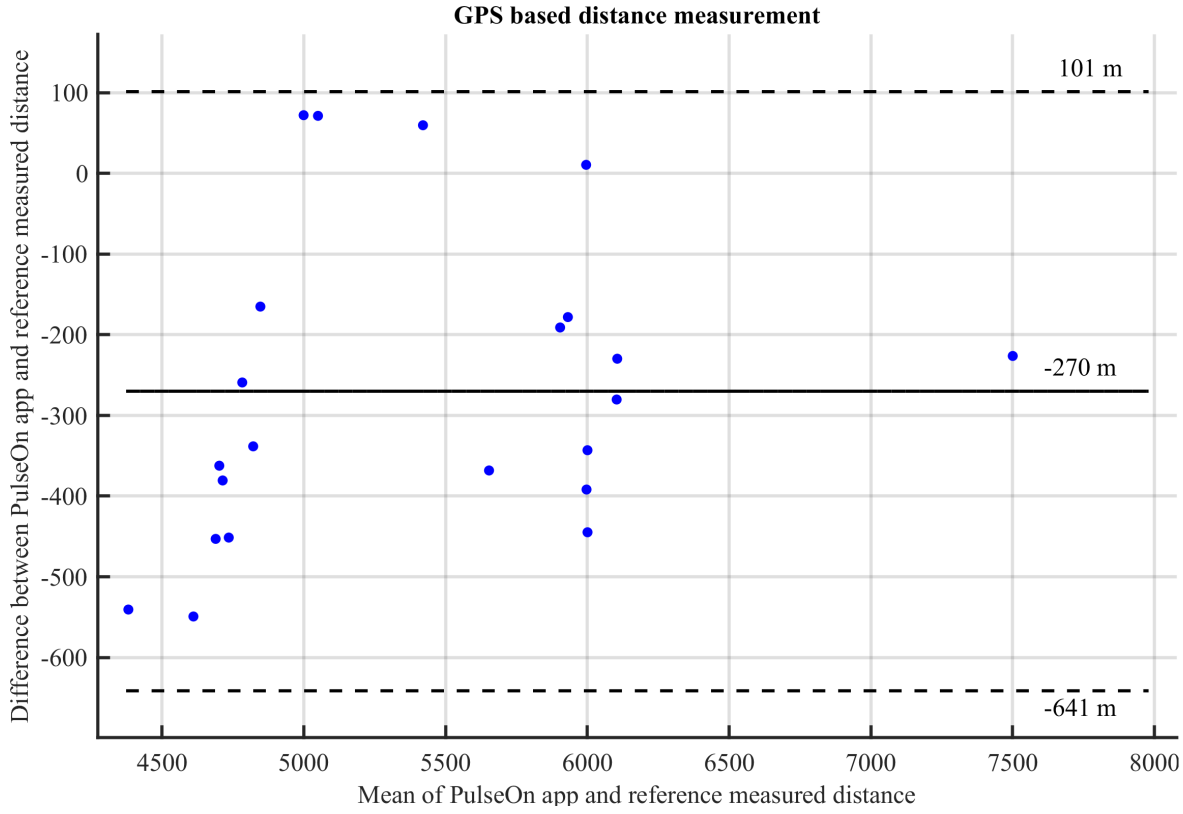


Figure 3. Bland−Altman plot comparing the phone GPS distance measured by the PulseOn app and a reference tracker distance estimation during outdoor running protocol (solid horizontal line: bias, dashed lines: 95% confidence limits of agreement).



Maximal Oxygen Intake Capacity Estimation

VO_{2Max} estimated with a wrist PPG device and phone GPS data with a PulseOn app was close to VO_{2Max} measured during maximum exercise tests in laboratory conditions (Tables 3 and 4). VO_{2Max} estimates were slightly underestimated with the submaximal test with the PulseOn app with a MAPE of 5.2% (4.7% for males and 5.8% for females), when measured maximum HR was used in the estimation. The distance estimated by a phone GPS was underestimated on average by 5.0% (−270m) (Figure 3). However, this error did not correlate with the VO_{2Max} error. When an age-based maximum HR estimate was used, the error slightly increased (Table 4). There was no statistically significant difference between the estimates when the measured maximum HR was used in the estimation. Figure 4 presents the Bland-Altman plot of the VO_{2Max} estimates, which shows a tendency towards larger errors with lower VO_{2Max} values.

Table 2. Accuracy of wrist optical heart rate device during treadmill running up to maximum speed.

Activity	Reliability, %	Accuracy, %
Rest when standing	96.9	97.1
Ramp-up running	95.3	98.3
Entire protocol	95.4	98.1

Table 3. Maximal oxygen uptake (VO_{2Max}) estimated from optical heart rate data and based on measured maximum heart rate value.

Performance metric	All (N=24)	Male (n=13)	Female (n=11)
Bias (ml ·kg^{-1}·min^{-1})	−1.07	−1.28	−0.82
SD[a] (ml ·kg^{-1} ·min^{-1})	2.75	2.42	3.19
MAE[b] (ml ·kg^{-1}·min^{-1})	2.39	2.29	2.51
MAPE[c]	5.2	4.7	5.8
Statistical test (*P* value)	.06(W[d])	.08(T[e])	.42(T[e])
Correlation coefficient	ρ=0.86, (*P*<.01)(Sp[f])	*r*=.77, (*P*<.01) (Pe[g])	*r*=.69, (*P*<.05) (Pe[g])

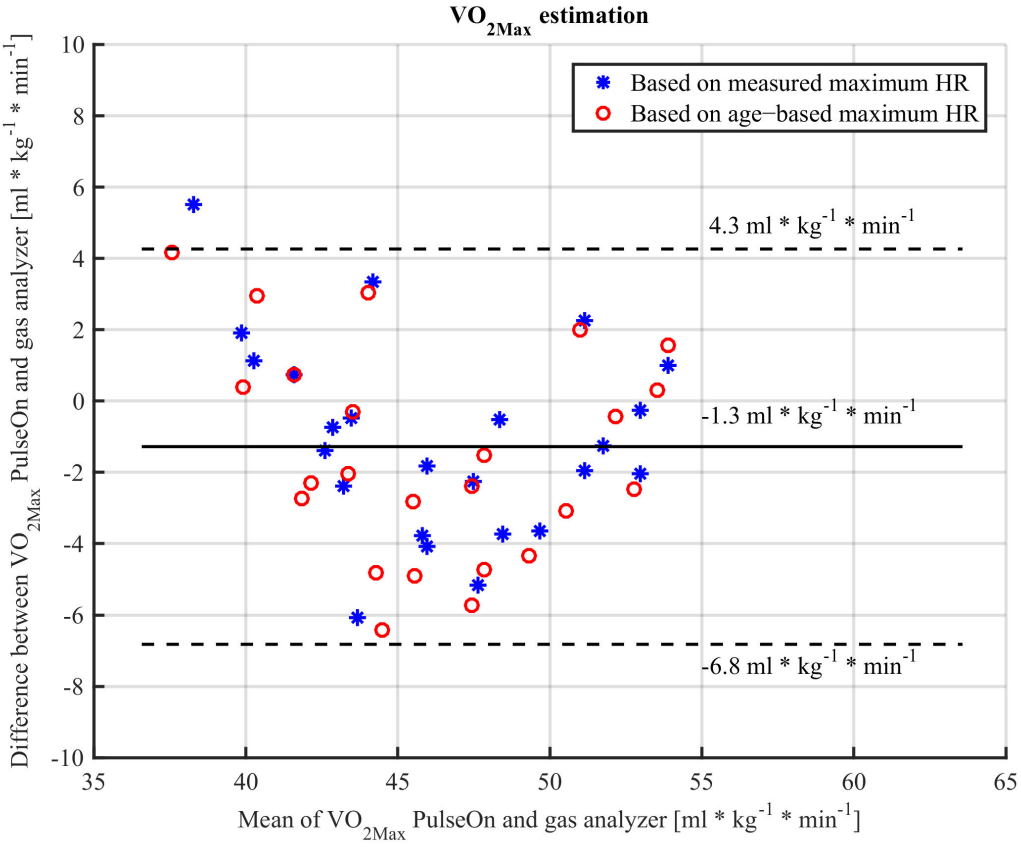
[a]SD: Standard deviation.
[b]MAE: Mean absolute error.
[c]MAPE: Mean absolute percentage error.
[d]W: Wilcoxon test.
[e]T: Paired *t* test.
[f]Sp: Spearman correlation coefficient.
[g]Pe: Pearson correlation coefficient.

Table 4. Maximal oxygen uptake (VO_{2Max}) estimated from optical heart rate data and based on an age-based maximum heart rate estimate.

Performance metric	All (N=24)	Male (n=13)	Female (n=11)
Bias (ml ·kg^{-1} ·min^{-1})	−1.49	−1.52	−1.46
SD[a] (ml ·kg^{-1} ·min^{-1})	2.95	2.70	3.35
MAE[b] (ml ·kg^{-1} ·min^{-1})	2.76	2.58	2.96
MAPE[c], %	5.9	5.2	6.8
Statistical test (*P* value)	.03(W[d])	.07(T[e])	.18(T[e])
Correlation coefficient	ρ=0.87, (*P*<.01)(Sp[f])	*r*=.73, (*P*<.01) (Pe[g])	*r*=.63, (*P*<.05) (Pe[g])

[a]SD: Standard deviation.
[b]MAE: Mean absolute error.
[c]MAPE: Mean absolute percentage error.
[d]W: Wilcoxon test.
[e]T: Paired *t* test.
[f]Sp: Spearman correlation coefficient.
[g]Pe: Pearson correlation coefficient.

Figure 4. Bland−Altman plot of VO2Max estimates from the PulseOn app (wrist PPG device + phone GPS) during a submaximal exercise test versus gas analyzer based estimate during maximal exercise tests−dots represent data when age−based maximum HR is used for an estimation, while an asterix represents estimations based on true measured maximum HR (solid horizontal line: bias, dashed lines: 95% confidence limits of agreement).



Energy Expenditure

Data from one male subject was excluded from the EE estimation analysis due do failure in respiratory gas analysis data acquisition, and results are reported for the remaining 23 subjects. Error in the EE estimation was lower (MAPE 6.7%) in the higher intensity exercise (above the aerobic threshold, but below the anaerobic threshold), but increased in lower intensities (Tables 5 and 6, and Figure 5). A wrist PPG device tended to underestimate the EE during treadmill running. The correlation with respiratory gas estimated EE was high (>.93) during higher intensity exercise, especially in females.

Table 5. Statistical error analysis of energy expenditure during light intensity.

Performance metric	All (N=23)	Male (n=12)	Female (n=11)
Bias (kcal)	−11.93	−14.24	−9.41
SD[a] (kcal)	13.99	16.45	10.95
MAE[b] (kcal)	13.05	15.28	10.65
MAPE[c], %	16.5	16.6	16.3
Statistical test (*P* value)	<.001 (W[d])	.01 (T[e])	.02 (T[e])
Correlation coefficient[e]	ρ=0.77, (*P*<.01) (Sp[f])	*r*=.88, (*P*<.01) (Pe[g])	*r*=.79, (*P*<.01) (Pe[g])

[a]SD: Standard deviation.

[b]MAE: Mean absolute error.

[c]MAPE: Mean absolute percentage error.

[d]W: Wilcoxon test.

[e]T: Paired *t* test.

[f]Sp: Spearman correlation coefficient.

[g]Pe: Pearson correlation coefficient.

Figure 5. Bland−Altman plot comparing an energy expenditure estimation with a wrist PPG device and gas analyzer during a maximum exercise test−the asterisk denotes data before the aerobic threshold, while dots represent data between aerobic and anaerobic thresholds (solid horizontal line: bias, dashed lines: 95% confidence limits of agreement).

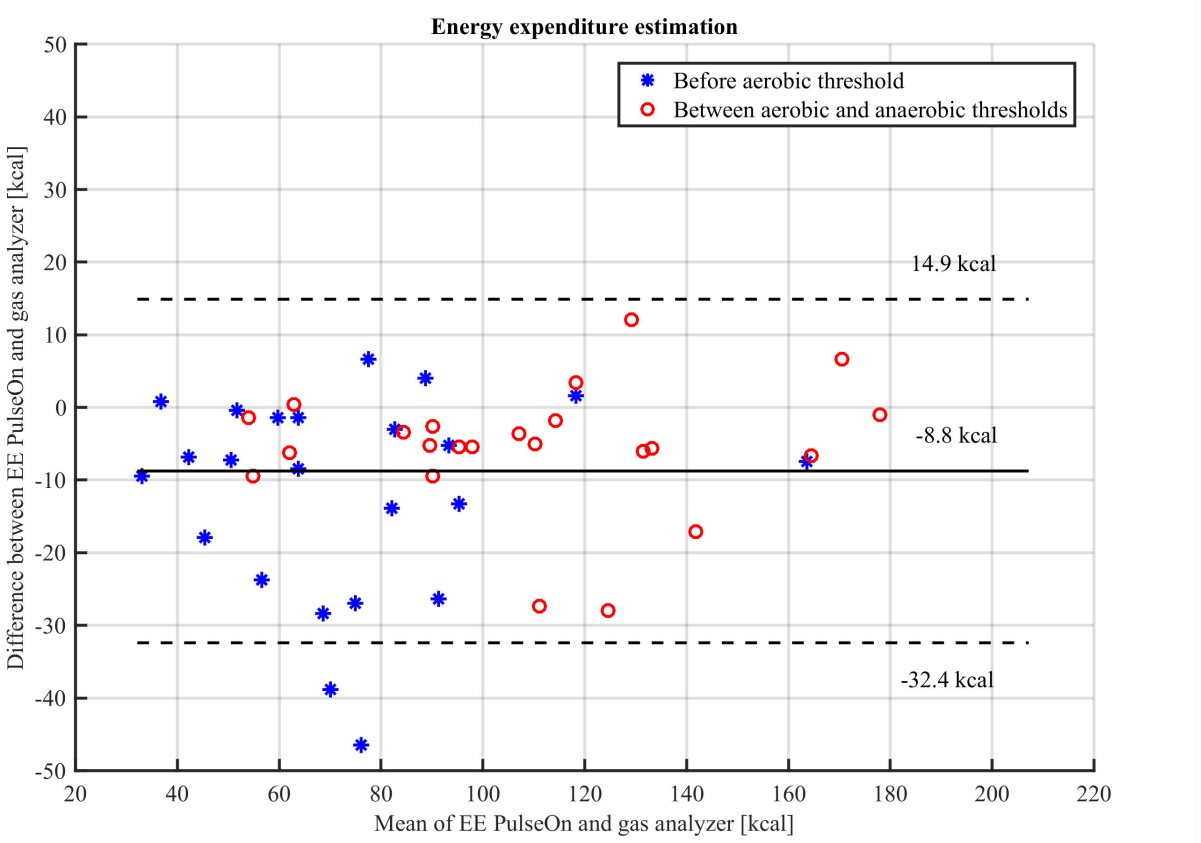


Table 6. Statistical error analysis of energy expenditure during medium heavy intensity.

Performance metric	All (N=23)	Male (n=12)	Female (n=11)
Bias (kcal)	−5.58	−6.78	−4.28
SD[a] (kcal)	9.00	12.24	3.10
MAE[b] (kcal)	7.52	10.43	4.34
MAPE[c], %	6.7	8.2	5.1
Statistical test (*P* value)	.007 (T[d])	.08(T[d])	.001(T[d])
Correlation coefficient	*r*=.97, (*P*<.01) (Pe[e])	*r*=.93, (*P*<.01) (Pe[e])	*r*=.99, (*P*<.01) (Pe[e])

[a]SD: Standard deviation.

[b]MAE: Mean absolute error.

[c]MAPE: Mean absolute percentage error.

[d]T: Paired *t* test.

[e]Pe: Pearson correlation coefficient.

Discussion

Principal Findings

We estimated HR, EE, and VO_{2Max} based on wrist PPG and phone GPS speed and evaluated their accuracy during running based on golden reference methods. OHR appeared to be accurate during running; the MAPE was 1.9% and reliability 95.4% during a maximal voluntary exercise test. This is well in line with the earlier results [4] and suggests that high-end consumer-grade OHR devices are capable of accurately monitoring HR during running, even up to a maximum HR.

The accuracy of more advanced parameters estimated from OHR is dependent, both on the accuracy of the OHR and on the validity of the analytical models. We used an HR-based estimation of the EE, and an HR and running speed-based estimation of the VO_{2Max} developed earlier by a company (Firstbeat, Jyväskylä, Finland), which is widely available in various sports products. EE estimation with this method has been validated earlier [8], suggesting a slight underestimation

of EE by 13% when a chest strap HR was used. In our study, the overall EE estimation accuracy is well in line with this. EE estimation was the most accurate during medium or hard intensity with a MAPE of 6.7% (males 8.2% and females 5.1%). During light intensity, the error increased to 16.5% (males 16.6% and females 16.3%). Differences in the EE estimation based on HR may be related to individual differences in the basic metabolism, a thermogenesis effect due to diet or metabolic effect, which affects the body mass ratio [25]. For comparison, 10.1-18.2% MAE has been reported for EE estimation by activity trackers [26]. The EE estimates based on OHR and indirect calorimetry had strong correlations for all (N=23) subjects during light intensity (ρ=0.77), while at a higher intensity their correlation was close to 1 (r=.97). The results are comparable to a similar study by Robertson et al [12], who used a chest strap HR with the same EE estimation method [18] and reported moderate (r=.57) to strong (r=.85) correlations during low and high intensity exercise, respectively. However, significant differences between different OHR devices have been reported. Recently, Wallen et al studied the EE estimation accuracy of four different OHR devices against indirect calorimetry and found only one device (Samsung Gear S) to have a strong correlation (r=.86) with the reference, while the other three devices exhibited only a weak correlation to reference EE [13]. Our results suggest that a wrist-worn OHR may offer a similar estimation of true EE during running to chest strap HR based methods when a high quality OHR device and proper physiological model are applied in EE estimation.

The level of fitness may be quantified by the estimation of VO_{2Max}. We used OHR and a mobile-based speed estimation to estimate VO_{2Max} during self-paced outdoor running in real and challenging outdoor conditions during winter in Finland. These conditions may be considered the "worst case" training conditions and, for example, the temperature difference may increase the observed estimation error for VO_{2Max}. The analytical method was based on the well-known HR versus speed relationship and on detecting the most reliable data periods for VO_{2Max} estimation during the exercise [20]. We compared this estimate with the golden standard of the VO_{2Max} estimation, that is, respiratory gas analysis acquired during a maximal voluntary exercise test in a sports laboratory. The results suggest that OHR and speed-based VO_{2Max} estimation during self-paced running are able to quite accurately estimate VO_{2Max}, even in these challenging outdoor conditions; we found a MAPE of 5.2% (males 4.7% and females 5.8%) for VO_{2Max} when an individually measured HR maximum was used in the estimation. When age-estimated maximum HR was used, the error increased slightly. A significant contribution to the inaccuracy originated from phone GPS tracking, which underestimated the distance by 5% on average and led to a corresponding underestimation of the VO_{2Max}. In addition, during the outdoor testing, there were challenging weather conditions (cold and winter), which posed challenges for PPG HR estimation because of potentially poor perfusion, increasing the potential error for the OHR during field conditions. These weather conditions may also have affected the real VO_{2Max}. Also, differences in running efficiency affect the correspondence between the running speed and the true physical load, and, hence, increase the error in HR and speed-based VO_{2Max} estimation. There was also a tendency for the OHR and speed-based analysis to overestimate the VO_{2Max} in individuals with a lower real VO_{2Max}. In summary, the results suggest that the method may be used to estimate VO_{2Max} relatively accurately during self-paced running, even in challenging outdoor conditions.

Limitations and Strengths

This study has several strengths, but also weaknesses. To our knowledge, this is the first study to report both EE and VO_{2Max} estimation accuracy, based on OHR data. We used a realistic or even challenging setting (self-paced outdoor running in winter) to estimate VO_{2Max}. This is a setting that can be applied by an ordinary user, and as such, the method can be directly applied by healthy users to estimate their fitness levels. We used the golden standard (gas analyzer and controlled sports laboratory with maximal voluntary exercise) as a reference for EE and VO_{2Max}. The main weakness of the study is that it had a relatively small study population; however, despite this, the results can be considered to be at least indicative. In addition, the outdoor tests were carried out in a challenging environment (winter, cold, and sometimes potentially slightly slippery roads), increasing the error of the outdoor VO_{2Max} estimation. On the other hand, this provides the worst case scenario, and the results were still within an acceptable error margin. Finally, the study included only one wrist OHR device, which limits the generalizability of the results. Only a single device was used for practical reasons—wearing several devices in both laboratory and outdoor conditions would have complicated the study implementation. The PulseOn device was chosen for the study because, at the time of data collection, to our knowledge, other available wrist OHR devices did not support estimation of VO_{2Max} together with accurate data logging capability. However, the results are not without generalizability. The applied VO_{2Max} and EE estimation algorithm [18,20] has been validated with a chest strap HR monitor [12], is commercially widely available, and could be applied with other accurate OHR devices as well. Hence, we do not consider the results of the study to be specific to applied wrist devices only, but to OHR technology in general.

Conclusions

We applied a commercially available OHR device to estimate HR, EE, and VO_{2Max} during running and evaluated their accuracy against golden standard methods. The results show that current high-end wrist OHR devices may provide accurate HR that can be compared with a chest strap HR, during running, up to a maximum HR. When combined with proper analytics, OHR may be used to quite accurately estimate EE, especially during moderate to medium heavy intensity activities. An estimation of VO_{2Max} during self-paced outdoor running using OHR and a mobile phone's GPS data and proper HR analytics also allows a relatively accurate estimation of a fitness level (VO_{2Max}). Wrist PPG devices accompanied by phone apps provide a reliable alternative for training monitoring in realistic conditions.

Acknowledgments

We thank the Varala Sports Institute for their support in the data collection and sports laboratory testing.

Conflicts of Interest

JP and IK are employees of PulseOn, Finland. PulseOn employed MU and JM during the period when the evaluation study was conducted.

References

1. El-Amrawy F, Nounou MI. Are currently available wearable devices for activity tracking and heart rate monitoring accurate, precise, and medically beneficial? Healthc Inform Res 2015 Oct;21(4):315-320 [FREE Full text] [doi: 10.4258/hir.2015.21.4.315] [Medline: 26618039]
2. Olenick A, Haile L, Dixon C. Validation of the Mio alpha heart rate monitor during graded exercise testing in trail runners. 2014 Nov Presented at: International Journal of Exercise Science: Conference Proceedings; 2014; Harrisburg, Pennsylvania.
3. Parak J, Korhonen I. Evaluation of wearable consumer heart rate monitors based on photopletysmography. In: Conf Proc IEEE Eng Med Biol Soc. 2014 Presented at: 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); August 26-30, 2014; Chicago, IL p. 3670-3673. [doi: 10.1109/EMBC.2014.6944419]
4. Delgado-Gonzalo R, Parak J, Tarniceriu A, Renevey P, Bertschi M, Korhonen I. Evaluation of accuracy and reliability of PulseOn optical heart rate monitoring device. In: Conf Proc IEEE Eng Med Biol Soc. 2015 Presented at: 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); August 25-29, 2015; Milan, Italy p. 430-433. [doi: 10.1109/EMBC.2015.7318391]
5. Spierer DK, Rosen Z, Litman LL, Fujii K. Validation of photoplethysmography as a method to detect heart rate during rest and exercise. J Med Eng Technol 2015;39(5):264-271. [doi: 10.3109/03091902.2015.1047536] [Medline: 26112379]
6. Stahl SE, An H, Dinkel DM, Noble JM, Lee J. How accurate are the wrist-based heart rate monitors during walking and running activities? are they accurate enough? BMJ Open Sport Exerc Med 2016 Apr 25;2(1):e000106. [doi: 10.1136/bmjsem-2015-000106]
7. Jo E, Lewis K, Directo D, Kim MJ, Dolezal BA. Validation of biofeedback wearables for photoplethysmographic heart rate tracking. J Sports Sci Med 2016 Sep;15(3):540-547 [FREE Full text] [Medline: 27803634]
8. Montgomery PG, Green DJ, Etxebarria N, Pyne DB, Saunders PU, Minahan CL. Validation of heart rate monitor-based predictions of oxygen uptake and energy expenditure. J Strength Cond Res 2009 Aug;23(5):1489-1495. [doi: 10.1519/JSC.0b013e3181a39277] [Medline: 19593221]
9. Keytel LR, Goedecke JH, Noakes TD, Hiiloskorpi H, Laukkanen R, van der Merwe L, et al. Prediction of energy expenditure from heart rate monitoring during submaximal exercise. J Sports Sci 2005 Mar;23(3):289-297. [doi: 10.1080/02640410470001730089] [Medline: 15966347]
10. Reis VM, den Tillaar RV, Marques MC. Higher precision of heart rate compared with VO2 to predict exercise intensity in endurance-trained runners. J Sports Sci Med 2011;10(1):164-168 [FREE Full text] [Medline: 24149310]
11. Charlot K, Cornolo J, Borne R, Brugniaux JV, Richalet J, Chapelot D, et al. Improvement of energy expenditure prediction from heart rate during running. Physiol Meas 2014 Feb;35(2):253-266. [doi: 10.1088/0967-3334/35/2/253] [Medline: 24434852]
12. Robertson AH, King K, Ritchie SD, Gauthier AP, Laurence M, Dorman SC. Validating the use of heart rate variability for estimating energy expenditure. International Journal of Human Movement and Sports Sciences 2015 Aug;3(2):19-26. [doi: 10.13189/saj.2015.030203]
13. Wallen MP, Gomersall SR, Keating SE, Wisløff U, Coombes JS. Accuracy of heart rate watches: implications for weight management. PLoS One 2016;11(5):e0154420 [FREE Full text] [doi: 10.1371/journal.pone.0154420] [Medline: 27232714]
14. Londeree BR. The use of laboratory test results with long distance runners. Sports Med 1986;3(3):201-213. [Medline: 3520749]
15. Nieman DC, editor. Submaximal laboratory tests. In: Exercise Testing and Prescription. New York: McGraw-Hill; 2011:52-59.
16. Leboeuf SF, Aumer ME, Kraus WE, Johnson JL, Duscha B. Earbud-based sensor for the assessment of energy expenditure, HR, and VO2max. Med Sci Sports Exerc 2014;46(5):1046-1052 [FREE Full text] [doi: 10.1249/MSS.0000000000000183] [Medline: 24743110]
17. Tanaka H, Monahan KD, Seals DR. Age-predicted maximal heart rate revisited. J Am Coll Cardiol 2001 Jan;37(1):153-156 [FREE Full text] [Medline: 11153730]
18. Firstbeat Technologies. Firstbeat. 2012 Mar. An energy expenditure estimation method based on heart rate measurement URL: https://assets.firstbeat.com/firstbeat/uploads/2015/10/white_paper_energy_expenditure_estimation.pdf [accessed 2017-01-18] [WebCite Cache ID 6nbZ10sXF]
19. McArdle W, Katch F, Katch V. Exercise physiologynergy, Nutrition and Human Performance. Baltimore, Philadephia: Lea & Febiger; 1996.

20. Firstbeat Technologies. Firstbeat. 2014 Nov. Automated fitness level (VO2max) estimation with heart rate and speed data URL: https://www.firstbeat.com/app/uploads/2015/10/white_paper_VO2Max_11-11-20142.pdf [accessed 2017-01-18] [WebCite Cache ID 6nbaC2uv7]
21. Weir JB. New methods for calculating metabolic rate with special reference to protein metabolism. J Physiol 1949 Aug;109(1-2):1-9 [FREE Full text] [Medline: 15394301]
22. Nummela A. Kestävyyssuorituskykyä selittävät tekijät [Endurance Perfomance Factors]. In: Keskinen K, Häkkinen K, Kalline M, editors. Kuntotestauksen käsikirja [Fitness testing handbook]. Helsinki: Liikuntatieteellinen Seura [Finnish Society of Sport Sciences]; 2007:51-59.
23. Jeukendrup AE, Wallis GA. Measurement of substrate oxidation during exercise by means of gas exchange measurements. Int J Sports Med 2005 Feb;26(Suppl 1):S28-S37. [doi: 10.1055/s-2004-830512] [Medline: 15702454]
24. Nummela A. Aerobisen kestävyyden suorat mittausmenetelmät [Direct aerobic endurance measurement methods]. In: Keskinen K, Häkkinen K, Kallinen M, editors. Kuntotestauksen käsikirja [Fitness testing handbook]. Helsinki: Liikuntatieteellinen Seura Finnish Society of Sport Sciences; 2007:64-78.
25. Donahoo WT, Levine JA, Melanson EL. Variability in energy expenditure and its components. Curr Opin Clin Nutr Metab Care 2004 Nov;7(6):599-605. [Medline: 15534426]
26. Evenson KR, Goto MM, Furberg RD. Systematic review of the validity and reliability of consumer-wearable activity trackers. Int J Behav Nutr Phys Act 2015;12(1):159 [FREE Full text] [doi: 10.1186/s12966-015-0314-1] [Medline: 26684758]

Abbreviations

HR: Heart rate
PPG: Photoplethysmography
OHR: Optical heart rate
BPM: Beats per minute
MAE: Mean absolute error
MAPE: Mean absolute percentage error
SD: Standard deviation
EE: Energy expenditure
VO2Max: maximal oxygen consumption

Edited by J Torous; submitted 02.02.17; peer-reviewed by M Nounou, Q Zhang; comments to author 22.03.17; revised version received 23.04.17; accepted 10.05.17; published 25.07.17

Please cite as:
Parak J, Uuskoski M, Machek J, Korhonen I
Estimating Heart Rate, Energy Expenditure, and Physical Performance With a Wrist Photoplethysmographic Device During Running
JMIR Mhealth Uhealth 2017;5(7):e97
URL: http://mhealth.jmir.org/2017/7/e97/
doi:10.2196/mhealth.7437
PMID:28743682

Tampereen teknillinen yliopisto
PL 527
33101 Tampere

Tampere University of Technology
P.O.B. 527
FI-33101 Tampere, Finland

ISBN 978-952-15-4210-7
ISSN 1459-2045