2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Accuracy of Wrist-Worn Photoplethysmography Devices at Measuring Heart Rate in the Laboratory and During Free-Living Activities.

Oonagh M. Giggins, *Member, IEEE*, Julie Doyle, Nisanth Sojan, Orla Moran, Daniel R Crabtree, Matthew Fraser, David J Muggeridge

***Abstract*—This study compared heart rate (HR) measurements taken from two wrist-worn devices; the Empatica E4 and the Apple Watch Series 5, to that taken from a Polar H10 chest strap. Ten healthy adult volunteers took part in a laboratory validation study and performed a treadmill exercise protocol. A single-subject validity study was also conducted to evaluate the accuracy of continuous HR measurements obtained during free-living activities. The participant wore both wrist devices, as well as the Polar H10 for 12-hours, as she continued her habitual daily activities. The key findings of the laboratory study were that the Apple Watch was accurate at assessing HR compared to the Polar H10 with Mean Absolute Percentage Error (MAPE) values < 5% during treadmill exercise. The accuracy of the E4 however was generally poor with MAPE values > 15%. Findings from the single-subject validity study indicate that the Apple Watch produces accurate measurements of HR, whereas the E4 device overestimated HR, except for during the more strenuous activities undertaken where HR was underestimated.**

***Clinical Relevance*— The Apple Watch has acceptable accuracy in measuring HR during treadmill exercise and during free-living activities in healthy adult volunteers.**

I. Introduction

Wearable physical activity monitors are becoming an increasingly popular means of tracking physical activity and have been shown to help people to modify and increase their physical activity levels [1]. These activity monitors use a variety of sensors and are typically incorporated into wrist-worn devices meaning they are particularly unobtrusive and easy to use. Many wrist-worn devices quantify physical activity by measurement of heart rate (HR) using a technique called photoplethysmography (PPG). Measuring HR allows for the intensity of exercise to be monitored, which is important during exercise for safety reasons and also to ensure optimal training intensities are achieved.

The potential for using wrist-worn devices, which incorporate PPG sensors has garnered considerable attention in healthcare. Some potential applications include use for screening/diagnostics purposes, in lifestyle monitoring and in therapeutic monitoring. However, if the information delivered is not accurate, the usefulness of these devices is limited. Many studies have been conducted examining the validity of HR measurements derived from wrist-worn PPG devices [2]–[4]. A recent systematic review found that these devices have acceptable validity at rest and during activities such as treadmill walking and running, however during cycling and resistance type exercise, measurements were not as accurate [5].

As new devices and new technologies emerge, it is important to test and validate these devices before they can be implemented clinically. Devices should also be evaluated in settings appropriate for their intended use [6]. While establishing the validity of devices in a controlled laboratory environment is important, it is also important to examine the validity of devices in naturalistic environments and during activities of daily living (ADL) where movements are more variable and sporadic.

This study examines two wrist-worn devices, namely the Empatica E4 (Empatica Inc., Cambridge, MA, USA), and the Apple Watch Series 5 (Apple Inc., Cupertino, CA, USA). These devices were selected for evaluation in this current study as they required testing prior to being used in other research being conducted by the authors. The Apple Watch has already been evaluated in a number of studies [5]. The E4 device has also been evaluated, albeit not as extensively as the Apple Watch, with the findings indicating that the E4 becomes more inaccurate with increasing activity [7], [8]. This study seeks to build upon this current literature base and aims to assess the validity of the HR measurements obtained from both devices in a laboratory setting and in the real-world.

II. Methodology

A cross-sectional laboratory-based validation study was conducted to examine the accuracy of HR measurements obtained from the Apple Watch and the E4 during treadmill exercise. In addition, a single-subject real-world validation study was conducted to determine the accuracy of continuous HR measurements produced by both devices under free-living conditions. The protocol was approved by the Health and

* This research is part of the ECME project which has been funded by the EU's INTERREG VA programme, managed by the Special EU Programmes Body (SEUPB).

O.M.G., J.D., N.S. and O.M are with NetwellCASALA, Dundalk Institute of Technology, Dundalk, Co. Louth, Republic of Ireland (corresponding author phone: +353-42-9370200 ext. 2114; e-mail: Oonagh.giggins@dkit.ie). D.R.C and M.F are with the Centre for Health Science, Division of Biomedical Sciences, University of the Highlands and Islands, Old Perth Road, Inverness IV2 3JH, Scotland, UK. D.M.J is with the School of Applied Sciences, Edinburgh Napier University, Edinburgh, UK.

Science Ethics Committee in Dundalk Institute of Technology (DkIT), and all participants provided written informed consent.

A. Laboratory Validation

A sample of convenience of healthy volunteers, aged 18-45 years participated. Exclusion criteria included any cardio-metabolic conditions, pacemakers, and any drug treatment/medical conditions that could interfere with HR measurements. Demographic and anthropometric data, and Fitzpatrick skin tone measurements [9] were recorded. The E4 (firmware version 3.1.0.7124) was positioned on the right wrist and the Apple Watch was worn on the left wrist (firmware version 6.0). The E4 device was paired with the E4 real-time application on an Android smartphone for data collection, while HR data were recorded from the Apple Watch using the Workout application on an Apple iPhone. A Polar H10 chest strap (Polar Electro, Kempele, Finland) was used as the criterion measure of HR. The Polar H10 was paired with the Polar Beat application for data acquisition. The treadmill protocol consisted of three five-minute stages; walking at 4km/h, jogging at 7km/h and running at 10km/h [10]. Each five minute stage of the protocol was interspersed with five minutes of rest sitting on a chair, and HR recovery was monitored for five minutes following the completion of the protocol.

B. Real-World Validation

A healthy female volunteer (age: 33 years; height: 169cm; weight: 61kg; Fitzpatrick skin tone type II, right hand dominant), who participated in the laboratory study, took part in the real-world validation. Recordings commenced at 9:00am and continued for a 12-hour period. The devices were worn as described in the laboratory study. The E4 logged and stored data to its internal memory. To simulate real-world usage, the workout application was not utilized on the Apple Watch, except for when it was used to record a run completed by the participant (35-minute outdoor run completed on a track). A digital notebook was used to document all activities undertaken, and the start and end times associated with each. Activities were classified as; sitting (any periods of sitting including quiet sitting, sitting while eating meals, sitting working at desk, typing etc.), standing, walking (purposive walking of more than 1 minute duration), running, or ADLs (cooking, cleaning, dressing etc.). The transition periods between the recorded activities were not removed from the data as per previous work [11].

C. Data Processing

Following data collection, the E4 data were synchronized with the E4 Connect web application and the Polar Beat data were synchronized with the Polar Flow web application. Data were downloaded from both web applications for analysis. The Apple Workout data were synced with the Apple Health application on the paired iPhone, and exported from Apple Health for analysis. Both the Polar H10 and the E4 device collected second-by-second HR data. Apple reports that the Apple Watch measures HR approximately every 10 minutes or continuously during a workout. However, on review of the data extracted, it appears to sample HR at a variable frequency (mean (SD) sampling rate = 0.19 (0.02) Hz) while using the workout application. For the laboratory validation, data from all devices were time-aligned and split according to the three five-minute stages of the exercise protocol. The average HR for each of these five-minute stages was then calculated for each device. Data collected from all devices during the real-world validation were also time-aligned and split according to the five activity domains. All HR data points collected were included in the analysis and comparisons between devices and the criterion were performed for each matched timestamp.

D. Statistical Analysis

Descriptive statistics of the mean and SD were used to summarize the data. Laboratory data were tested for normality using the Shapiro-Wilk test with a p-value of < 0.05 considered statistically significant. The difference between each devices' HR measure and the Polar H10 determined HR during each stage of the laboratory treadmill protocol and for each activity domain during the real-world validation was calculated. Agreement between each device and the Polar H10 was analyzed using the Bland–Altman method [12], where the mean bias, SD, and the upper and lower limits of agreement (LOA) were calculated. Mean absolute percentage error (MAPE) values were calculated as the average absolute value of the errors of each device relative to the Polar H10 determined measurement of HR. Spearman's rank correlation coefficient (r_s) was used to determine the strength of relationship and a significant correlation was determined if the p-value was < 0.05. Correlation coefficients were interpreted as follows: 0.9-1.0 = strong, 0.8-0.89 = moderately strong, 0.7-0.79 = moderate, 0.6-0.69 = moderately weak, <0.59 = weak.

III. Results

A. Laboratory Validation

Eleven recreationally active participants enrolled in this study and ten participants (five male, five female; age: 30.4 ± 8.0 years; height: 174.0 ± 9.4 cm; weight: 74.7 ± 12.1 kg). Fitzpatrick skin tone: type I n=2, type II n=2, type IIII n=3, type IV n=2, type V n=1) completed the protocol. Due to partial device failure, Apple Watch data are missing for four participants during the 10km/h stage of the exercise protocol and for one participant during the 7km/h stage of the protocol. The cases with missing data were omitted and all remaining data were analyzed. The results of the Shapiro Wilk test indicated normal distribution of the data. Correlation coefficient, mean bias, 95% LOA and MAPE for both devices during each stage of the laboratory treadmill protocol are indicated in Table 1. Across all stages of the protocol, the Apple Watch demonstrated a small mean bias. The Apple Watch demonstrated significant ($p < 0.001$) strong correlations with the Polar H10 during the 4km/h stage and the 7km/h stage. The E4 exhibited weak correlations during all stages. The E4 overestimated HR during the 4km/h stage and grossly underestimated HR during both the 7km/h and 10km/h stage of the protocol. The Apple Watch achieved an error rate of < 5% during all three exercise stages.

B. Real-World Validation

The E4 device collected second-by-second HR data, resulting in a total of 43,200 HR observations during the 12-hour data collection period. The Polar H10 device also collected data at a frequency of 1Hz, however data collection was erroneously paused for a brief period on the Polar Beat application, resulting in a total of 43,166 Polar HR observations. The Apple Watch collected a total 568 HR observations over the 12-hour period. Only two HR observations were recorded on the Apple Watch during the standing and walking activities domains, therefore descriptive statistics are only presented for these two domains. Correlation coefficient, mean bias, 95% LOA and MAPE for both devices during each of the activity domains are presented in Table 2. For both the running and sitting activity domain, the Apple Watch demonstrated a small mean bias and significant ($p < 0.05$) moderately-strong to strong correlations with the Polar H10, while the E4 device exhibited weak correlations across all activity domains. The E4 device overestimated HR during each activity domain except for during the running activity where it underestimated HR with an error rate of 28%.

IV. Discussion

The findings of this study indicate that the Apple Watch provides the most accurate measure of HR relative to the criterion in both the laboratory and during free-living activities. Across all intensities of treadmill exercise, strong correlations ($r = 0.856 – 0.962$) were observed between the Apple Watch and Polar H10, together with MAPE values <5%. MAPE values of ±10% are typically interpreted as acceptable error rate [11]. In addition, strong correlations ($r = 0.876 – 0.986$) were also observed during outdoor running and for the sitting activity domain in the real-world validation study, with MAPE values within the acceptable range. For the ADL domain however the MAPE was greater than the ±10% threshold and this may be explained by the erratic arm movements that sometimes occur during ADLs. Overall, the Apple Watch may be viewed as accurate in measuring HR in healthy adult volunteers during treadmill exercise and during free-living activities, and these findings are in line with the evidence presented in the literature [5].

TABLE 1 RESULTS FOR EACH DEVICE DURING EACH STAGE OF THE LABORATORY TREADMILL PROTOCOL

	Polar	E4	Apple
4km/h Stage	(n=10)	(n=10)	(n=10)
Mean ± SD (BPM)	89.8 ±14.2	98.4 ± 10.0	91.3 ± 12.9
Mean Bias ± SD (BPM)	-	-8.6 ± 16.6	-1.6 ± 4.6
95% LOA (upper, lower)	-	24.0, -41.1	7.3, -10.4
MAPE	-	15.1	3.9
r_s	-	0.430	0.976***
7km/h	(n=10)	(n=10)	(n=9)
Mean ± SD (BPM)	126.3 ±19.8	110.8 ±20.2	127.2 ±25.8
Mean Bias ± SD (BPM)	-	15.5 ± 23.9	-2.3 ± 8.3
95% LOA (upper, lower)	-	62.3, -31.3	14.0, -18.5
MAPE	-	15.8	4.6
r_s	-	0.442	0.967***
10km/h	(n=10)	(n=10)	(n=6)
Mean ± SD (BPM)	150.8 ± 8.7	106.6 ±24.8	156.5 ± 9.5
Mean Bias ± SD (BPM)	-	44.3 ± 31.1	1.3 ± 5.6
95% LOA (upper, lower)	-	105.3, -16.8	12.3, -9.6
MAPE	-	28.4	2.6
r_s	-	-0.079	0.771

Spearman's rank correlations (r_s) between the Polar H10 and the Empatica.E4 and the Apple Watch. *** correlation is significant at $p < 0.001$

TABLE 2 RESULTS FOR EACH DEVICE DURING ALL FIVE ACTIVITY DOMAINS

	Polar	E4	Apple
Activities of daily living			
Observations n	20247	20281	57
Mean ± SD (BPM)	65.0 ± 9.9	81.5 ± 5.1	65.4 ± 3.6
Mean Bias ± SD (BPM)	-	-16.4 ±18.1	-5.8 ± 13.6
95% LOA (upper, lower)	-	19.1,-51.9	20.9, -32.5
MAPE	-	-28.4	-11.3
r_s	-	-0.001	0.462***
Running			
Observations n	2460	2460	406
Mean ± SD (BPM)	126.7 ±36.1	81.5 ± 7.7	132.1 ±41.0
Mean Bias ± SD (BPM)	-	45.2 ± 34.7	-5.5 ± 7.2
95% LOA (upper, lower)	-	113.2, -22.8	8.6, -19.6
MAPE	-	27.8	-4.0
r_s	-	0.030	0.693***
Sitting			
Observations n	19018	19018	101
Mean ± SD (BPM)	55.8 ± 9.5	73.4 ± 14.8	68.6 ± 18.8
Mean Bias ± SD (BPM)	-	-17.6 ±17.1	-7.5 ± 9.6
95% LOA (upper, lower)	-	15.9, -51.1	15.3, -22.3
MAPE	-	-35.0	-7.5
r_s	-	0.025**	0.743***
Walking			
Observations n	1081	1081	2
Mean ± SD (BPM)	83.4 ± 16.8	93.1 ± 19.4	64.5 ± 20.5
Mean Bias ± SD (BPM)	-	-9.7 ± 31.1	-
95% LOA (upper, lower)	-	51.3,-70.7	-
MAPE	-	-19.1	-
r_s	-	-0.262***	-
Standing			
Observations n	360	360	2
Mean ± SD (BPM)	53.1 ± 7.7	70.4 ± 8.5	54.0 ± 5.7
Mean Bias ± SD (BPM)	-	-17.3 ± 12.1	-
95% LOA (upper, lower)	-	6.4, -41.0	-
MAPE	-	-35.3	-
r_s	-	-0.540***	-

Spearman's rank correlations (r_s) between the Polar H10 and the Empatica.E4 and the Apple Watch. Correlation is significant at $^{**}p < 0.01$, $^{***}p < 0.001$

The E4 device does not accurately measure HR during treadmill walking and like previous investigations, demonstrated a substantial increase in measurement error as the intensity of exercise increased [7]. Additionally, during free-living activities, the E4 device appears to produce inaccurate measurements of HR. Mean bias increases substantially as the level of activity increases, with the highest error rate observed during the running activity domain. During non-strenuous activities such as the sitting and ADL domains, the MAPE was also observed to be outside the acceptable range. Overall, it appears that the HR signal obtained from the E4 device is greatly compromised by motion artefact caused by increasing arm movement, as in previous research [13], [14].

This study has a number of limitations which should be acknowledged. Firstly, the laboratory study examined a small sample of healthy adult volunteers who exercised in a controlled environment on a treadmill. Previous work has shown that the type of activity influences the accuracy of HR measurements obtained from wrist-worn devices [8]. Future studies should include a range of different activities. The sample included was a convenience sample of healthy, younger individuals who engaged in regular exercise and were within a healthy body mass index range. Therefore the results cannot be generalized to older adults or to individuals of other

body sizes. There are suggestions that differences in skin tone can have an impact on PPG derived measurement of HR [5], [11]. While skin tone was recorded in this study using the Fitzpatrick Skin Scale [9], differences in HR measurements between those of different skin tones was not examined. Future validation studies should investigate these differences more thoroughly. Great care was taken in our investigations to ensure that the devices were placed correctly and according to the manufacturers' recommendations. However as some participants had larger or smaller wrists, there may have been inconsistencies in device placement once the participant started moving.

These demographic variables and other potential confounding factors were controlled for in the single-subject study. This real-world validation study allowed for the collection of continuous HR data in a naturalistic environment as the participant engaged with her day-to-day activities, therefore providing a robust validation of devices. While the single-subject design in itself is a limitation, this approach has wide appeal and has been extensively used in the fields of psychology, education, and human behavior [15] and there is growing support for combining the data generated from such single-subject studies [16]. This work should therefore be replicated across a number of subjects and the results meta-analyzed.

In both investigations, the Polar H10 was used to provide the criterion measure of HR. While many other studies have also utilized chest straps as the reference measurement of HR [17][18], they do have a degree of error, and a 12-lead ECG should be considered the gold standard. Another shortcoming of this work is the differences in the sampling rate used by the devices. Both the E4 device and Polar H10 collected data at a frequency of 1Hz, however the Apple Watch acquired HR data at a variable frequency rate. This resulted in a substantially smaller number of HR observations for each activity domain from the Apple Watch with only two observations recorded for both the standing and walking activity domains. Comparisons between the Apple Watch and the Polar H10 were performed for each matched timestamp, and may have produced more favorable results for the Apple Watch.

V. Conclusion

This study examined the accuracy of HR measurements obtained from the Empatica E4 and the Apple Watch in healthy adult volunteers. This study found that the Apple Watch produced accurate HR measures during treadmill exercise. In addition, continuous measurements of HR produced by the Apple Watch during free-living activities were also within an acceptable error range. The E4 device did not produce accurate measurements of HR and appears to be compromised by motion artefact. Further evaluations of the Apple Watch are required to determine the accuracy of HR measures in clinical cohorts.

References

[1] J. B. Wang *et al.*, "Wearable Sensor/Device (Fitbit One) and SMS Text-Messaging Prompts to Increase Physical Activity in Overweight and Obese Adults: A Randomized Controlled Trial," *Telemed. e-Health*, vol. 21, no. 10, pp. 782–792, Jun. 2015.

[2] B. D. Boudreaux *et al.*, "Validity of Wearable Activity Monitors during Cycling and Resistance Exercise.," *Med. Sci. Sports Exerc.*, vol. 50, no. 3, pp. 624–633, Mar. 2018.

[3] D. J. Muggeridge *et al.*, "Measurement of Heart Rate Using the Polar OH1 and Fitbit Charge 3 Wearable Devices in Healthy Adults During Light, Moderate, Vigorous, and Sprint-Based Exercise: Validation Study," *JMIR Mhealth Uhealth*, 2021.

[4] A. Khushhal *et al.*, "Validity and Reliability of the Apple Watch for Measuring Heart Rate During Exercise," *Sport. Med. Int. Open*, vol. 1, no. 06, pp. E206--E211, 2017.

[5] Y. Zhang, R. G. Weaver, B. Armstrong, S. Burkart, S. Zhang, and M. W. Beets, "Validity of Wrist-Worn photoplethysmography devices to measure heart rate: A systematic review and meta-analysis," *J. Sports Sci.*, vol. 38, no. 17, pp. 2021–2034, Sep. 2020.

[6] F. Sartor, G. Papini, L. G. E. Cox, and J. Cleland, "Methodological Shortcomings of Wrist-Worn Heart Rate Monitors Validations," *J. Med. Internet Res.*, vol. 20, no. 7, p. e10108, Jul. 2018.

[7] B. Bent, B. A. Goldstein, W. A. Kibbe, and J. P. Dunn, "Investigating sources of inaccuracy in wearable optical heart rate sensors," *npj Digit. Med.*, vol. 3, no. 1, p. 18, 2020.

[8] L. Barrios, P. Oldrati, S. Santini, and A. Lutterotti, "Evaluating the Accuracy of Heart Rate Sensors Based on Photoplethysmography for In-the-Wild Analysis," in *Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare*, 2019, pp. 251–261.

[9] T. B. Fitzpatrick, "The validity and practicality of sun-reactive skin types I through VI," *Arch Dermatol*, vol. 124, 1988.

[10] A. Khushhal *et al.*, "Validity and Reliability of the Apple Watch for Measuring Heart Rate During Exercise," *Sport. Med. Int. Open*, vol. 1, no. 06, pp. E206–E211, 2017.

[11] B. W. Nelson and N. B. Allen, "Accuracy of Consumer Wearable Heart Rate Measurement During an Ecologically Valid 24-Hour Period: Intraindividual Validation Study.," *JMIR mHealth uHealth*, vol. 7, no. 3, p. e10828, Mar. 2019.

[12] J. M. Bland and D. G. Altman, "Statistical methods for assessing agreement between two methods of clinical measurement.," *Lancet (London, England)*, vol. 1, no. 8476, pp. 307–310, Feb. 1986.

[13] E. A. Thomson *et al.*, "Heart rate measures from the Apple Watch, Fitbit Charge HR 2, and electrocardiogram across different exercise intensities.," *J. Sports Sci.*, vol. 37, no. 12, pp. 1411–1419, Jun. 2019.

[14] J. Pietilä *et al.*, "Evaluation of the accuracy and reliability for photoplethysmography based heart rate and beat-to-beat detection during daily activities BT - EMBEC & NBC 2017," 2018, pp. 145–148.

[15] J. D. Smith, "Single-case experimental designs: a systematic review of published research and current standards.," *Psychol. Methods*, vol. 17, no. 4, pp. 510–550, Dec. 2012.

[16] E. O. Lillie, B. Patay, J. Diamant, B. Issell, E. J. Topol, and N. J. Schork, "The n-of-1 clinical trial: the ultimate strategy for individualizing medicine?," *Per. Med.*, vol. 8, no. 2, pp. 161–173, Mar. 2011.

[17] M. Etiwy *et al.*, "Accuracy of wearable heart rate monitors in cardiac rehabilitation.," *Cardiovasc. Diagn. Ther.*, vol. 9, no. 3, pp. 262–271, Jun. 2019.

[18] G. Abt, J. Bray, and A. C. Benson, "The validity and inter-device variability of the Apple Watch for measuring maximal heart rate.," *J. Sports Sci.*, vol. 36, no. 13, pp. 1447–1452, Jul. 2018.