Reliability Assessment of New and Updated Consumer-Grade Activity and Heart Rate Monitors

Salome Oniani
Faculty of Informatics and Control Systems
Georgian Technical University
Tbilisi, Georgia
E-mail: s.oniani@gtu.ge

Sandra I. Woolley
School of Computing and Mathematics
Keele University
Staffordshire, UK
E-mail: s.i.woolley@keele.ac.uk

Ivan Miguel Pires[*,**] and Nuno M. Garcia[*]
[*] Instituto de Telecomunicações,
Universidade da Beira Interior
Covilhã, Portugal
[**] Altranportugal
Lisbon, Portugal
E-mails: impires@it.ubi.pt, ngarcia@di.ubi.pt

Tim Collins
School of Engineering
Manchester Metropolitan University
Manchester, UK
E-mail: t.collins@mmu.ac.uk

Sean Ledger and Anand Pandyan
School of Health and Rehabilitation
Keele University
Staffordshire, UK
E-mails: s.j.ledger@keele.ac.uk, a.d.pandyan@keele.ac.uk

***Abstract*— The aim of this paper is to address the need for reliability assessments of new and updated consumer-grade activity and heart rate monitoring devices. This issue is central to the use of these sensor devices and it is particularly important in their medical and assisted living application. Using an example lightweight empirical approach, experimental results for heart rate acquisitions from Garmin VivoSmart 3 (v4.10) smartwatch monitors are presented and analyzed. The reliability issues of optically-acquired heart rates, especially during periods of activity, are demonstrated and discussed. In conclusion, the paper recommends the empirical assessment of new and updated activity monitors, the sharing of this data and the use of version information across the literature.**

Keywords- wearable sensing; activity monitoring; ambulatory heart rate, inter-instrument reliability.

I. INTRODUCTION

Consumer-grade wearable monitoring devices are used across a spectrum of health, well-being and behavioral studies, as well as clinical trials. For example, the U.S. Library of Medicine ClinicalTrials.gov database reports nearly 200 "Completed" to "Not yet recruiting" trials involving Fitbit devices (search accessed 01/05/2018). However, the manufacturers of these devices are generally very clear regarding the intended applications and suitability of their devices, and do not make misleading clinical claims. For example, Garmin Vivosmart "Important Safety and Product Information" [1] advises that the device is for "*recreational purposes and not for medical purposes"* and that *"inherent limitations"* may *"cause some heart rate readings to be inaccurate"*, similarly, Fitbit device "Important Safety and Product Information" declares that the device is *"not a medical device"* and *"accuracy of Fitbit devices is not intended to match medical devices or scientific measurement devices"* [2]. Given that these devices *are* being used in clinical applications, and with future clinical applications anticipated [3], it is important that device reliability is assessed.

In terms of meeting user expectations, it is noteworthy that, at the time of writing, Fitbit's motion to dismiss a class action has been denied. The complaint alleged *"gross inaccuracies and recording failures"* [4] because *"products frequently fail to record any heart rate at all or provide highly inaccurate readings, with discrepancies of up to 75 bpm"* [5]. Indeed, ambulatory heart rate acquisition from optical sensors is known to be very challenging [6]. One of the main challenges is the range of severe interference effects caused by movement [7][8]. Optical heart rate signals can also be affected by skin color [9] and aging [10]. Yet, optical heart rate acquisition remains a desirable alternative to chest strap electrocardiogram (ECG) monitoring in consumer-level activity monitors, where comfortability, ease-of-use and low cost are prioritized.

After selection of an activity monitor model based on recorded parameters, study requirements and deployment needs [11], the calibration and validation of wearable monitors [12][13] can be onerous. Best practice requires a substantial time and resource investment for researchers to calibrate and validate sufficiently large numbers of their devices with a large and diverse cohort of representative

 ISBN: 978-1-61208-660-6

users performing a range of anticipated activities. At the same time, commercial monitors can frequently and automatically update both software and firmware that can alter device function, data collection and data reporting, potentially compromising previous validation. But, of course, manufacturers are under no obligation to report the detail of their proprietary algorithms or the specifics of version changes.

Devices that have the same model name, but operate with different software and firmware versions, are distinct devices; they should not be treated as identical devices. Ideally, devices would be clearly differentiated in the literature with data for manufacturer, model *and* version data. While there may be limited (if any) opportunity for researchers to reversion commercial device software to repeat published experiments, the provision of version information would, at least, limit the potential for incorrect aggregations of data for devices that operate with different software and firmware versions.

A number of studies have reported on the validity of different monitoring device models. For example, Fokkema et al. [14] reported on the step count validity and reliability of ten different activity trackers. Thirty-one healthy participants performed 30-minute treadmill walking activities while wearing ten activity trackers. The research concluded that, in general, consumer activity trackers perform better at an average (4.8 km/h) and vigorous (6.4 km/h) walking speed than at slower walking speeds.

In another study, Wahl et al. [15] evaluated the validity of eleven wearable monitoring devices for step count, distance and Energy Expenditure (EE) with participants walking and running at different speeds. The study reported results with the commonly used metrics: Mean Absolute Percentage Error (MAPE) and IntraClass Correlation (ICC) showing that most devices, except Bodymedia Sensewear, Polar Loop, and Beurer AS80 models, had good validity (low MAPE, high ICC) for step count. However, for distance, all devices had low ICC (<0.1) and high MAPE (up to 50%), indicating poor validity. The measurement of EE was acceptable for Garmin, Fitbit and Withings devices (comprising Garmin Vivofit; Garmin Vivosmart; Garmin Vivoactive; Garmin Forerunner 920XT; Fitbit Charge; Fitbit Charge HR; Withings Pulse Ox Hip; Withings Pulse Ox Wrist) which had low-to-moderate MAPEs. The Bodymedia Sensewear, Polar Loop, and Beurer AS80 devices had high MAPEs (up to 56%) for all test conditions.

There is a growing number of similar studies that compare different recordings from different models of consumer activity monitors. However, across this literature, and in reviews of this literature [16], it is common practice to provide version data for the software used for statistical analyses of device performance, but it is not common practice to report version information for the devices themselves. As an example of device ambiguity, a reference to "Garmin Vivosmart" could refer to either Garmin Vivosmart 3 or Garmin Vivosmart HR. The date of a given publication might help disambiguate the model variant but will not help identify the version. The Vivosmart HR had 14 versions from 2.10 to 4.30 over approximately 30 months (each update comprising between 1 and 11 items, such as, *"improved calculation of Intensity Minutes"* and *"Various other improvements"*) [17]. At the time of writing, the Garmin Vivosmart 3 (v4.10) is the latest of 9 versions.

In Section II of this paper, a lightweight approach for device assessment is presented using the Garmin Vivosmart 3 smartwatch as an example device; the results of an experimental assessment are presented in Section III. Recommendations for device assessment are discussed in Section IV and, conclusions and recommendations for further work are summarized in Section V.

II. METHOD AND MATERIALS

Four Garmin Vivosmart 3 smartwatches (all versioned SW v4.10 throughout the data acquisitions during May 2018) were worn, as shown in Figure 1, by four healthy researcher participants, P1-P4 outlined in Table I, during the treadmill walking activities summarized in Table II. The walking speeds: slow, moderate, fast and vigorous, were selected based on reports in the literature [18][19] and were performed on an h/p/cosmos Pulsar treadmill. To support reproducibility [20], we report further details about materials in the appendix.

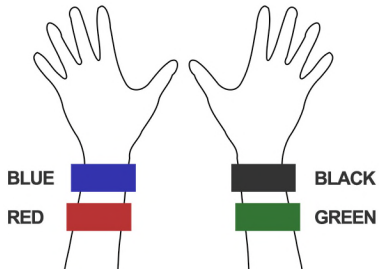

Figure 1. Activity monitor positions (color-coded for reference).

TABLE I. PARTICIPANT SUMMARY

Participant	Age (yrs)	Gender	Height (m)	Weight (kg)	BMI
P1	25	Female	1.69	58	20.03
P2	54	Female	1.62	65	24.7
P3	47	Male	1.75	70	22.8
P4	28	Male	1.70	76	26.2

TABLE II. THE WALKING ACTIVITY SCHEDULE

Time (minutes)	20	20	20	20
Activity	Slow walking (2.4 km/h)	Moderate walking (3.2 km/h)	Fast Walking (4.8 km/h)	Vigorous walking (6.4 km/h)

All participants reported regularly partaking brisk-intensive exercise outside largely sedentary academic/working roles. Participant 1 was ambidextrous. All other participants were right-handed. (Ethical approval for "Health Technology Assessment and Data Analytics", ERP2329" was obtained from Keele University.)

 ISBN: 978-1-61208-660-6

The slow walking activity was prefaced by two minutes of standing with arms down. Pulse readings were taken from a Polar H10 chest strap ECG monitor at 1-minute intervals throughout the activity.

Data (from the logged Garmin .FIT files) was downloaded from the watches after each activity and converted into .CSV formats and imported into Excel. Dates and times were converted from the Garmin 16- and 32-bit timestamps used in the .FIT file [21] into standard Excel date-time serial numbers.

Mean Absolute Percentage Error (MAPE) and the IntraClass Correlation (ICC) [22] were used to compare the heart rate recordings from the watches with the baseline ECG device. Step counts were also acquired and analyzed but, due to limitations of space, are not reported here.

III. RESULTS

Figure 2 shows the heart rate recordings for P1-P4 from the treadmill walking activities. Variability in recorded values can be seen at both slower and faster walking speeds and, notably, differs between participants. For analysis of the acquired data we calculated the MAPE (compared with the ECG chest strap reference) and ICC values listed in Table III. As shown, treadmill acquisitions for participants P2 and P3 produced higher MAPEs (including MAPEs over 10%: the level often taken as the upper bound of "acceptable" errors) and lower ICCs. This could, in part, be attributed to the increased age of participants P2 and P3 compared to P1 and P4. As shown in Figure 2, for P2 there were some abnormally low but sparse heart rate recordings from the "blue" device and, to a lesser extent, the "red" device. For P3, the "blue" device recorded decreasing heart rates when the actual heart rate increased during the vigorous walking activity. This produced a near zero ICC.

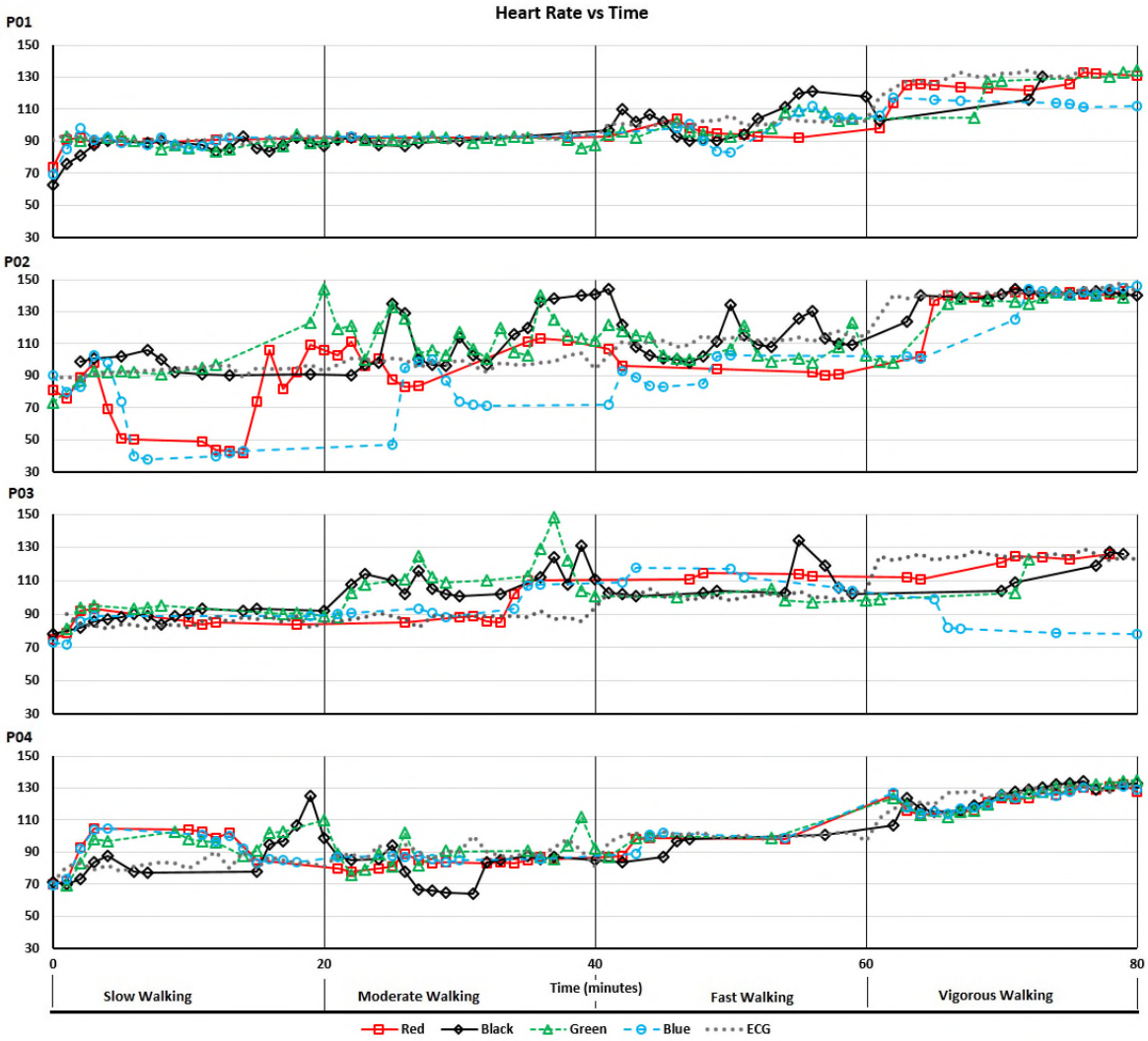

Figure 2. Heart rate recordings acquired during treadmill walking activities.

TABLE III. VALUES OF MAPE AND ICC FROM TREADMILL WALKING ACTIVITIES

Participant	*Black*		*Blue*		*Green*		*Red*	
ID	*MAPE*	*ICC*	*MAPE*	*ICC*	*MAPE*	*ICC*	*MAPE*	*ICC*
P1	7.08%	0.68	7.13%	0.71	4.34%	0.81	5.62%	0.90
P2	9.60%	0.69	15.55%	0.67	11.94%	0.58	13.42%	0.71
P3	13.00%	0.47	14.00%	0.02	16.00%	0.19	9.00%	0.84
P4	8.69%	0.84	6.14%	0.91	8.04%	0.86	7.57%	0.89

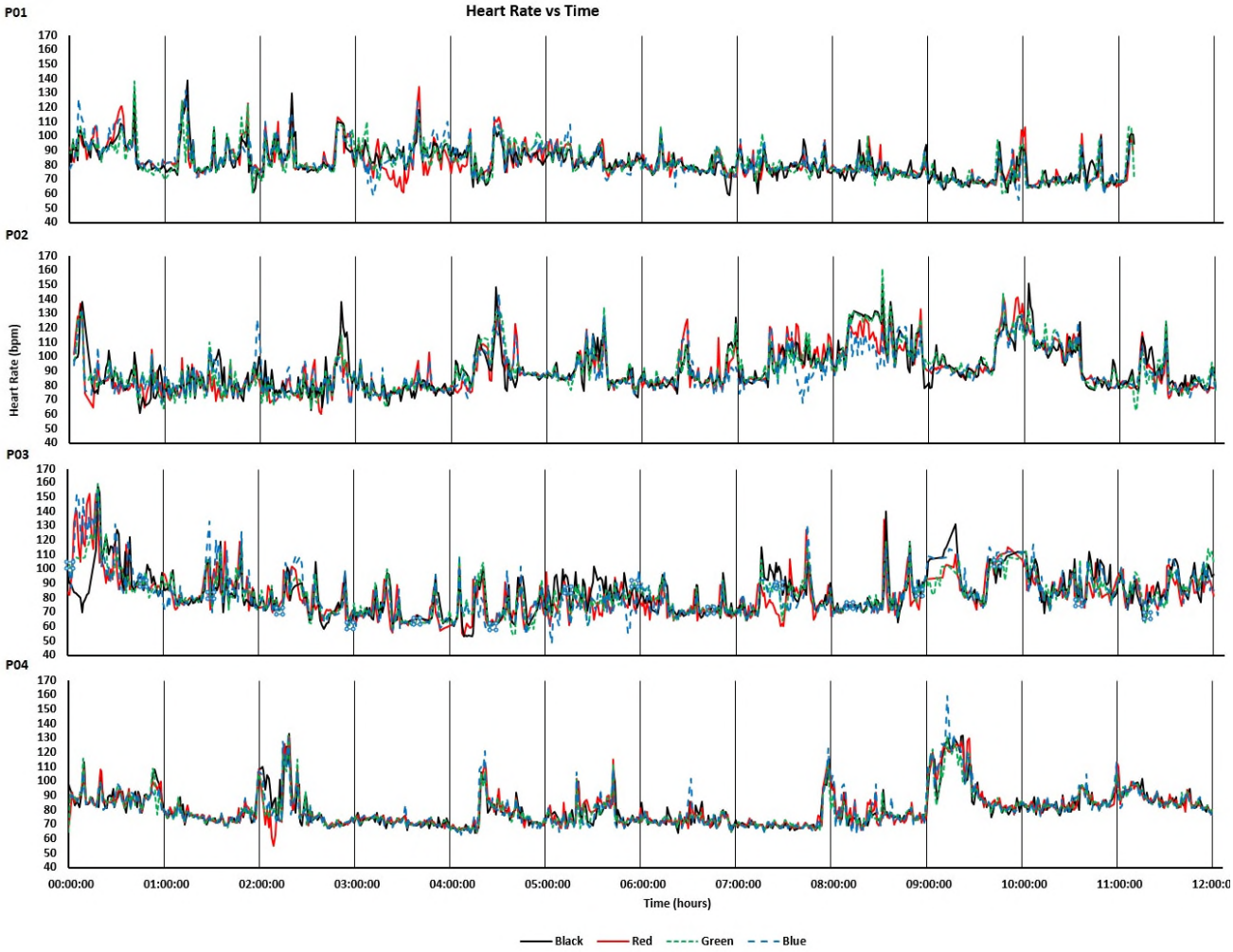

Figure 3. Heart rate recordings acquired during 12-hr everyday living.

The devices were also worn by participants for 12-hour periods during uncontrolled everyday activities. The recorded heart rates are shown in Figure 3. Intraclass correlations and confidence intervals for treadmill walking and 12-hr use are plotted, respectively, in Figures 4 and 5. As anticipated, these indicated poor performance during the treadmill activity. However, as shown in Figure 5, the devices performed more consistently during the prolonged acquisitions of activities of everyday living, when activity levels were generally lower on average.

IV. DISCUSSION

The lightweight assessment approach exemplified here is not, and could not be, prescriptive. A useful approach must incorporate participants and activities that have relevance to the intended study; otherwise, it would have little value. It is also important to ensure that the duration of activities is sufficient for devices to record enough data. We established 20-minute durations empirically for each treadmill walking speed by monitoring the frequency of logged readings and expanding the window to ensure several readings would be logged for each speed. For other devices where, for example, per-minute records are available, the activity duration could be reduced.

 ISBN: 978-1-61208-660-6

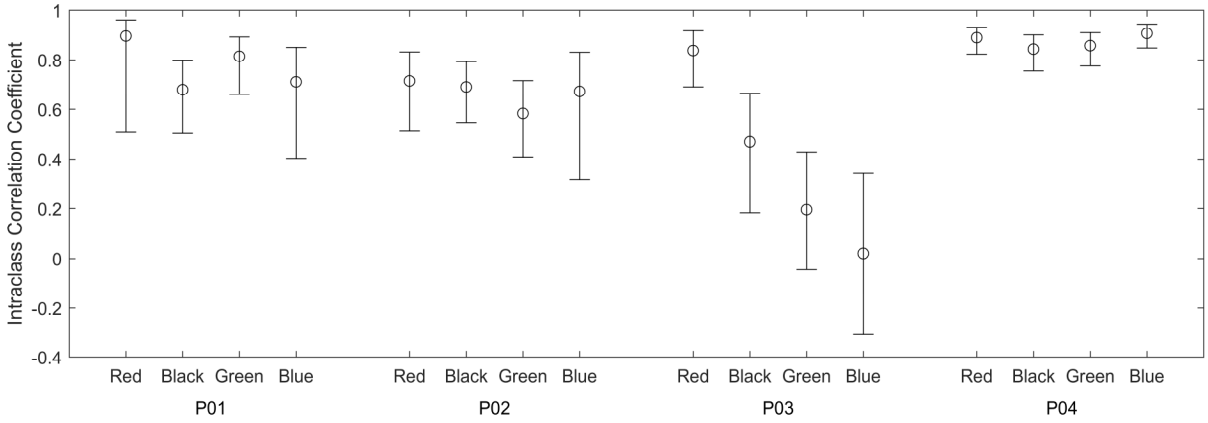

Figure 4. ICC for each device compared with ECG chest strap baseline recordings with 90% confidence intervals for treadmill activities.

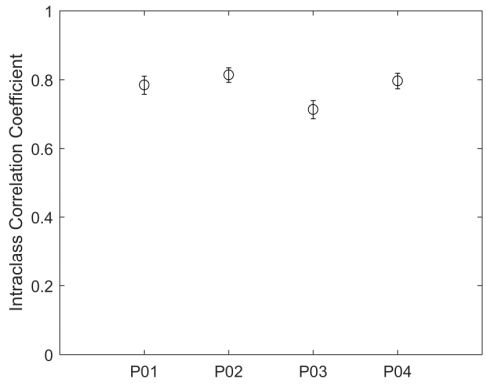

Figure 5. Inter-instrument ICC values for 12-hrs everyday living.

Of course, a comprehensive reliability assessment would be preferable to the approach outlined here. Similarly, this lightweight empirical approach is preferable to no assessment at all or reliance on outdated, irrelevant or unreproducible reports in the literature. Of the several limitations of the presented approach, there was, intentionally, a small number of participants, a limited sample of unrepeated activities and there were no reference recordings for the 12-hr everyday activity. (Reference readings from finger-worn pulse oximeters were attempted, but the devices repeatedly failed to maintain accurate readings). However, with just four participants and two activity acquisitions, we were able to quickly and simply obtain an insight into the reliability of the devices *at their current version*, have an appreciation of their limitations and, also, a degree of confidence regarding their potential for study acquisitions.

V. CONCLUSION AND FUTURE WORK

There is much scope for further work to improve reproducibility across the activity monitoring domain and to assist researchers evaluate and re-evaluate new and updated devices. We have demonstrated an empirical approach to device assessment that provides an example lightweight assessment that is not onerous and could easily be repeated as and when devices are updated.

Despite issues associated with reliable optical heart rate acquired from the wrist during activity, we might hope that future and updated consumer devices would i) be better at identifying erroneous values and avoid reporting them and ii) be better at correctly estimating values. It would be unwise to assume all device upgrades will necessarily result in improved device performance in all aspects, however, future research into sensor positioning, sensor array configurations, multi-sensor fusions and advanced signal processing techniques could significantly contribute to improved sensing reliability.

The U.S. Food and Drug Administration has established a new "Digital Health Software Precertification (Pre-Cert) Program" [23] that aspires toward a more agile approach to digital health technology regulation. It recognizes the *"iterative characteristics"* of new consumer devices [24]. In addition, the Consumer Technology Association recently defined CTA-2065; a new protocol to test and validate the accuracy of heart rate monitoring devices under the conditions of everyday living – from dynamic indoor cycling to sedentary lifestyles. We recommend that there is also some means to enable and encourage the sharing of version-by-version device reliability assessment data between manufacturer/s, users and researchers.

In a systematic review of consumer-wearable activity trackers, Everson et al. [16], recommend that *"future studies on the measurement properties of the trackers should be sure to initialize the tracker properly and indicate in the publication how this was done so others can replicate the process. Providing the specific tracker type, date purchased, and date tested would also be important."* We additionally recommend that full device details, including software and firmware versions, are reported in the literature.

ACKNOWLEDGEMENT

The authors wish to thank Professor Fiona Polack, Software and Systems Engineering Research, Keele

University for her valuable input and support in resourcing this work. The authors also thank Professor Barbara Kitchenham for her advice on protocol design and statistics.

The authors also wish to acknowledge contributions from FCT project UID/EEA/50008/2013 and COST Actions IC1303 (AAPELE – Architectures, Algorithms and Protocols for Enhanced Living Environments) and IC1401 (Memristors - Devices, Models, Circuits, Systems and Applications (MemoCiS)).

APPENDIX

The further material details were as follows:

Garmin Vivosmart 3 software/firmware versions: SW: v4.10; TSC: v1.10; SNS: v5.90. Devices were initialized according to the arm worn and all data was taken directly from logged .FIT files. Devices were purchased on 9th March 2018 and acquisitions made during May 2018. Their serial numbers were as follows: Black – 560185378, Red – 560185383, Blue – 560640435, Green – 560639717.

The treadmill was an h/p/cosmos Pulsar treadmill, h/p/cosmos Sports & Medical Gmbh, Nussdorf-Traunstein, Germany. (cos100420b; ID: X239W80479043; OP19: 0319 1139)

Polar H10 chest heart rate monitor (FCC ID: INW1W; Model: 1W; IC: 6248A-1W; SN: C7301W0726005; ID: 14C00425; Firmware: 2.1.9 and data acquired via Polar Beat 2.5.3.

REFERENCES

[1] Garmin Vivosmart 2016, “Important safety and product information,” Instruction Leaflet supplied with Vivosmart 3 (v4.10), l90-02068-0l_0D

[2] Fitbit, “Important safety and product information,” Last Updated March 20, 2017, [Online]. Available from: https://www.fitbit.com/uk/legal/safety-instructions 2018.06.02

[3] M. M. Baig, H. GholamHosseini, A. A. Moqeem, F. Mirza, and M. Lindén, “A systematic review of wearable patient monitoring systems–current challenges and opportunities for clinical adoption,” Journal of Medical Systems, vol. 41(7): 115, pp. 1-9, 2017.

[4] Business Wire, “Federal court denies Fitbit's motion to dismiss class action lawsuit alleging gross inaccuracies and recording failures in PurePulse™ heart rate monitors, June 05, 2018: [Online]. Available from: https://www.businesswire.com/news/home/20180605006652/en/Federal-Court-Denies-Fitbits-Motion-Dismiss-Class 2018.06.02

[5] Lieff Cabraser Civil Justice Blog, June 5, 2018, [Online]. Available from: https://www.lieffcabraser.com/2018/06/federal-court-denies-fitbit-motion-to-dismiss-class-action-lawsuit-inaccuracies-purepulse-heart-rate-monitors/ 2018.06.07

[6] M. Lang, “Beyond Fitbit: A critical appraisal of optical heart rate monitoring wearables and apps, their current limitations and legal implications,” Albany Law Journal of Science & Technology 28(1), pp. 39-72, 2017.

[7] Z. Zhang, “Heart rate monitoring from wrist-type photoplethysmographic (PPG) signals during intensive physical exercise,” In Signal and Information Processing (GlobalSIP), IEEE Global Conference on, pp. 698-702, December 2014.

[8] Z. Zhang, Z. Pi, and B. Liu, “TROIKA: A general framework for heart rate monitoring using wrist-type photoplethysmographic signals during intensive physical exercise,” IEEE Transactions on Biomedical Engineering, vol. 62(2), pp. 522-531, 2015.

[9] W. T. Cecil, K. J. Thorpe, E. E. Fibuch, and G. F.Tuohy, “A clinical evaluation of the accuracy of the Nellcor N-100 and Ohmeda 3700 pulse oximeters,” Journal of Clinical Monitoring, vol. 4(1), pp. 31-36, 1988.

[10] K. S. Hong, K. T. Park, and J. M. Ahn, “Aging index using photoplethysmography for a healthcare device: comparison with brachial-ankle pulse wave velocity,” Healthcare Informatics Research, vol. 21(1), pp. 30-34, 2015.

[11] T. Collins, S. Aldred, S. I. Woolley, and S. Rai, “Addressing the deployment challenges of health monitoring devices for a dementia study,” In Proceedings of the 5th EAI International Conference on Wireless Mobile Communication and Healthcare, pp. 202-205, 2015.

[12] D. R. Bassett Jr, A. V. Rowlands, and S. G. Trost, “Calibration and validation of wearable monitors”, Medicine and Science in Sports and Exercise, 44(1 Suppl 1), p.S32, 2012.

[13] P. Freedson, H. R. Bowles, R. Troiano, and W. Haskell,. “Assessment of physical activity using wearable monitors: recommendations for monitor calibration and use in the field,” Medicine and Science in Sports and Exercise, vol. 44(1 Suppl 1):S1-S4, pp. 1-6, 2012.

[14] T. Fokkema, T. J. Kooiman, W. P. Krijnen, C. P. Van Der Schans, and M. De Groot, “Reliability and validity of ten consumer activity trackers depend on walking speed,” Medicine and Science in Sports and Exercise, 49(4), pp. 793-800, 2017.

[15] Y. Wahl, P. Düking, A. Droszez, P. Wahl, and J. Mester, “Criterion-validity of commercially available physical activity tracker to estimate step count, covered distance and energy expenditure during sports conditions,” Frontiers in Physiology, vol. 8:725, pp. 1-12, 2017.

[16] K. R. Evenson, M. M. Goto, and R. D. Furberg, "Systematic review of the validity and reliability of consumer-wearable activity trackers," International Journal of Behavioral Nutrition and Physical Activity, vol. 12(1):159, pp. 1-22, 2015.

[17] Garmin 2018, “Updates & Downloads: vivosmart HR software - version 4.30 as of March 7, 2018,” [Online]. Available from: https://www8.garmin.com/support/download_details.jsp?id=9527 2018.05.30

[18] P. M. Grant, P.M. Dall., S. I. Mitchell, and M. H.Granat, “Activity-monitor accuracy in measuring step number and cadence in community-dwelling older adults,” Journal of Aging and Physical Activity, 16(2), pp. 201-214, 2008.

[19] J. Takacs, et al. “Validation of the Fitbit One activity monitor device during treadmill walking,” Journal of Science and Medicine in Sport, vol. 17(5), pp. 496-500, 2014.

[20] S. Krishnamurthi and J. Vitek, “The real software crisis: Repeatability as a core value,” Communications of the ACM, vol. 58(3), pp. 34-36, 2015.

[21] Garmin, 2018. FIT Software Development Kit (version 20.56.00), [Online]. Available from: https://www.thisisant.com/resources/fit 2018.06.02

[22] K. O. McGraw and S. P. Wong, “Forming inferences about some intraclass correlation coefficients,” Psychological methods, vol. 1(1), p. 30-46, 1996.

[23] U.S. Food and Drug Administration “Digital health software precertification (pre-cert) program,” [Online]. Available from: https://www.fda.gov/MedicalDevices/DigitalHealth/DigitalHealthPreCertProgram/ucm567265.htm 2018.06.13

[24] “CTA announces standard to improve heart rate monitoring in wearables,” May 2, 2018, [Online]. Available from: https://www.cta.tech/News/Press-Releases/2018/May/CTA-Announces-Standard-to-Improve-Heart-Rate-Monit.aspx 2018.06.1

 ISBN: 978-1-61208-660-6