The validity and inter-device variability of the Apple Watch™ for measuring maximal heart rate

Running title

Apple Watch maximal heart rate validity and variability

Grant Abt[1], James Bray[1], Amanda Clare Benson[2]

[1]School of Life Sciences, The University of Hull, U.K;

[2]Department of Health and Medical Sciences, Swinburne University of Technology, Australia

Correspondence

Grant Abt Ph.D.
School of Life Sciences
The University of Hull
Kingston upon Hull, HU6 7RX
United Kingdom
Email: g.abt@hull.ac.uk

James Bray Ph.D.
School of Life Sciences
The University of Hull
Kingston upon Hull, HU6 7RX
United Kingdom
Email: j.bray@hull.ac.uk

Amanda Benson Ph.D.
Department of Health and Medical Sciences

Swinburne University of Technology

Hawthorn, 3122

Australia

Email: abenson@swin.edu.au

Key words

Intensity, Validity, Reliability, Technology.

Word count

2418

Abstract

Maximal heart rate (HR_{max}) is a fundamental measure used in exercise prescription. The Apple Watch™ measures heart rate yet the validity and inter-device variability of the device for measuring HR_{max} are unknown. Fifteen participants completed a maximal oxygen uptake test while wearing an Apple Watch™ on each wrist. Criterion HR_{max} was measured using a Polar T31™ chest strap. There were good to very good correlations between the watches and criterion (left: r = 0.87 [90%CI: 0.67 to 0.95]; right: r = 0.98 [90%CI: 0.94 to 0.99]). Standardised mean bias for the left and right watches compared to the criterion were 0.14 (90%CI: -0.12 to 0.39; trivial) and 0.04 (90%CI: -0.07 to 0.15; trivial). Standardised typical error of the estimate for the left and right watches compared to the criterion were 0.51 (90%CI: 0.38 to 0.80; moderate) and 0.22 (90%CI: 0.16 to 0.34; small). Inter-device standardised typical error was 0.46 (90%CI: 0.36 to 0.68; moderate), ICC = 0.84 (90%CI: 0.65 to 0.93). The Apple Watch™ has good to very good criterion validity for measuring HR_{max}, with no substantial under- or over-estimation. There were moderate and small prediction errors for the left and right watches. Inter-device variability in HR_{max} is moderate.

Introduction

Heart rate is often used to prescribe exercise intensity in both the general population and athletes. Heart rate reserve (HRR), of which maximal heart rate (HR_{max}) is a fundamental component, is the method recommended for setting the exercise intensity of a training session (Achten & Jeukendrup, 2003; Pescatello & American College of Sports Medicine, 2014). The maximal heart rate included in the HRR method can be obtained by direct measurement during a maximal exercise test or predicted using a variety of age-based formulae such as the commonly used 220 – age or the more precise 206.9 – (0.67 x age) (Gellish et al., 2007) which is now the age-predicted HR_{max} estimation formula recommended by the American College of Sports Medicine (Pescatello & American College of Sports Medicine, 2014). Unfortunately, age-based formulas have considerable prediction error (Gellish et al., 2007; Robergs & Landwehr, 2002) which means that direct measurement of HR_{max} is still the preferred option if possible.

Heart rate can be directly measured by palpation, commonly at the carotid or radial pulse, using an ECG (which is not readily accessible to the general public), a telemonitoring device, or more recently using photoplethysmography (PPG). Most commercial heart rate monitors (e.g. Polar™) detect the electrical signals from the heart, however, during free-living conditions wearing a chest strap for lengthy periods of time is not always feasible, desirable or comfortable. In contrast, PPG measures heart rate using optical sensors that detect changes in the volume of blood flow in the capillaries below the skin (Allen, 2007). PPG optical sensors shine light through the skin to enable the detection of changes in blood volume perfusion of microvascular tissue. These changes are analysed using computer-based pulse-wave analysis techniques to determine heart rate (Allen, 2007). The measurement of heart rate using an optical sensor placed directly on the skin therefore negates the need for the user to wear a chest strap. However, motion artefact, for example that which might be observed with movement of the sensor across the skin, can cause measurement error of up to 8% through ambiguous automated waveform labelling when compared to that of an ECG

(Allen, 2007). For quality measurement using PPG, reduced movement of the sensor on the skin is imperative (Hertzman, 1938).

The consumer ‘wearables’ market, which includes smartwatches, has grown considerably over the last few years with forward estimates placing the size of the market at over 200 million devices to be sold in 2020 (IDC, 2016). More specifically, the Apple Watch™, which includes a PPG sensor for measuring heart rate, is reported to have had sales of more than 12 million units in 2015, making it the world’s most popular smartwatch (Canalys, 2016). This growth in the wearables market and the popularity of smartwatches like the Apple Watch™ have considerable potential for the promotion and monitoring of physical activity and exercise. Moreover, monitoring heart rate via a smartwatch enables incrementally progressive exercise prescription to maximise health related benefits and the provision of instant user feedback to assist with safety, motivation and adherence (Lyons, Lewis, Mayrsohn, & Rowland, 2014; Pescatello & American College of Sports Medicine, 2014). For these reasons, the measurement of heart rate via a smartwatch needs to be both valid and reliable.

Despite the popularity of the Apple Watch™, there have been a limited number of studies examining its validity for measuring heart rate during a variety of submaximal activities (Wallen, Gomersall, Keating, Wisløff, & Coombes, 2016; Wang et al., 2016). Wallen et al. (2016) compared the validity of four devices for measuring HR (Apple Watch™, Fitbit Charge HR™, Samsung Gear S™, Mio ALPHA™), with the Apple Watch™ having the lowest mean difference (SD) (-1.3 (4.4) beats·min^{-1}) and limits of agreement (-9.9 to 7.3 beats·min^{-1}) compared with an ECG. However, HR was manually recorded during the submaximal activities and the process of how HR data were extracted from the four devices is not clearly explained. It is also unknown on which arm each of the four devices was worn and which two devices were tested together (Wallen et al., 2016). Although comparisons were made between the four devices in both studies they were not tested simultaneously (Wallen et al., 2016; Wang et al., 2016) and only two of the four devices were worn in the study by Wang et al. (2016), which may have caused other unaccounted for measurement error despite randomisation (Wallen et al., 2016; Wang et al., 2016) and counterbalance allocation (Wallen et al., 2016). Wang et al. (2016) examined the validity of the Apple Watch™ HR compared to an ECG and a Polar chest strap. Participants exercised on a motorised treadmill at 3.2 km·h^{-1}, 4.8 km·h^{-1}, 6.4 km·h^{-1}, 8 km·h^{-1}, and 9.6 km·h^{-1} for 3 min at each stage while wearing two of four wrist-worn devices (Fitbit Charge HR™, Apple Watch™, Mio Alpha™, and Basis Peak™). There was a correlation of r = 0.91 (95%CI: 0.88 to 0.93) between the Apple Watch™ and the ECG. The limits of agreement range from -27 to +29 beats·min^{-1} compared to the ECG. However, HR was only taken once manually at the end of each 3-min stage which is a substantial limitation. There was also no indication on which wrist the Apple Watch™ was worn.

As such, there are substantial limitations with the previous Apple Watch™ studies and no study has examined the validity of the Apple Watch™ for measuring HR during maximal intensity exercise. Therefore, our aims were to examine the concurrent criterion validity of the Apple Watch™ for measuring HR$_{max}$, and to examine the variability in HR$_{max}$ between two Apple Watches worn simultaneously on the left and right wrists.

Methods

Fifteen (8 male, 7 female) recreationally active participants (those meeting the minimum ACSM guidelines for physical activity) (mean (SD) age 32 (10) y; body mass 73.5 (14.8) kg; stature 175 (8) cm) were enrolled in the study. Following University Human Ethics Committee approval (approval number 1516076), participants provided written informed consent prior to having their cardiovascular disease risk assessed according to the ACSM risk stratification guidelines (Pescatello & American College of Sports Medicine, 2014). All participants were stratified as low risk. Participants were recruited from the local community and university student body via written promotional material or personal request. Based on the data from Wallen et al. (2016) and Wang et al. (2016) who reported correlations of 0.98 and 0.91 between the Apple Watch and the criterion measure, it is not unreasonable to select

0.7 as the smallest correlation worth detecting. Using the formula $32/ES^2$, a sample of eight is derived (Hopkins, 2007). The correlation of 0.7 is considered to be the same as a Cohen's *d* effect size of 2.

A single maximal oxygen uptake test was used to establish the validity and inter-device variability of the Apple Watch for measuring maximal heart rate. The criterion measure of maximal heart rate was considered to be that measured by a Polar T31™ heart rate monitor.

Nude body mass was measured to the nearest 0.1 kg using digital scales (WB-100MA Mark 3, Tanita Corporation, Tokyo, Japan). Prior to the measure being taken participants were asked if they had voided prior to attending the session. If not, they were instructed to do so. Participants were then instructed to remove all clothing. Two measurements of body mass were then taken and the mean used for further analysis. Stretch stature was measured using a wall-mounted stadiometer (Holtain Ltd, Dyfed, Wales, UK) and according to the methods of the International Society for the Advancement of Kinanthropometry (Norton et al., 2000).

Participants completed a single incremental maximal oxygen uptake test on a motorised treadmill (h/p/cosmos, Pulsar, Nussdorf-Traunstein, Germany) while wearing a first-generation Apple Watch™ (watchOS 2.0.1) on each wrist (right and left) and a Polar T31™ chest strap (Polar, Kempele, Finland). The protocol commenced at 3 $km{\cdot}h^{-1}$ at a 1% gradient and increased 0.5 $km{\cdot}h^{-1}$ in speed every 30 s. Participants continued the protocol until volitional fatigue. Oxygen uptake was measured continuously from expired air using an online breath-by-breath system (Cortex Metalyzer 3B, GmbH, Germany). The analyser was calibrated before each test using room air and known gas concentrations of O_2 and CO_2. Volume was calibrated using a 3 L syringe.

Heart rate data were recorded every 5 s on each watch using the 'Workout' app. The 'Workout' app automatically syncs exercise data to the 'Health' database on its paired iPhone after the completion of an exercise session. To retrieve this raw heart rate and sampling time data a bespoke iPhone app was used. The bespoke app was written in Xcode

7.2.1 using the language Swift 2.1 and utilising the methods provided by the HealthKit framework (Apple, Inc). Criterion heart rate were measured using a Polar T31™ chest strap interfaced with a metabolic cart. The highest 30 s mean heart rate from each of the three devices (Polar T31^{TM}, left Apple Watch, right Apple Watch) were used as the values for maximal heart rate. Additionally, age-predicted maximal heart rate were calculated using both the 220 - age and 206.9 - [0.67 x age] formulas (Gellish et al., 2007). Although no verification phase was conducted, based on established criteria (volitional exhaustion; RER > 1.15; plateau in oxygen consumption < 150 mL·min^{-1}) (Howley, Bassett, & Welch, 1995), all participants were judged to have reached maximal oxygen consumption and therefore by association, HR$_{max}$.

Data were log transformed prior to analysis to avoid bias resulting from non-uniformity of error. Differences in the mean heart rate between the criterion and Apple Watch™ are reported as Cohen's *d*, together with 90% confidence intervals. Apple Watch™ validity (N = 14; missing HR data were excluded on one occasion as the Polar T31™ monitor did not record heart rate data) is reported as a Pearson correlation (r), standardised mean bias, and standardised typical error of the estimate (Hopkins, 2015). The 95% limits of agreement were calculated to enable comparison with other studies. Inter-device variability (N = 15) is reported as the standardised typical error and intraclass correlation (ICC). Uncertainty is reported as a 90% confidence interval. All data were analysed using custom-designed Microsoft Excel spreadsheets (Hopkins, 2015).

Results

The mean (SD) maximal heart rate (HR$_{max}$) were 183 (12) and 182 (12) beats·min^{-1} for the left and right Apple Watch™, respectively (mean difference -1 beats·min^{-1} [90%CI: -4 to 2]). Mean (SD) HR$_{max}$, as measured by the Polar T31™ chest strap (criterion), was 180 (12) beats·min^{-1}. There was a good correlation between the left Apple Watch™ and the criterion and a very good correlation between the right Apple Watch™ and the criterion (Figure 1).

Standardised mean bias for the left and right Apple Watch™ compared to the criterion were 0.14 (90%CI: -0.12 to 0.39; trivial) and 0.04 (90%CI: -0.07 to 0.15; trivial). Standardised typical error of the estimate for the left and right Apple Watch™ compared to the criterion were 0.51 (90%CI: 0.38 to 0.80; moderate) and 0.22 (90%CI: 0.16 to 0.34; small). The mean bias and 95% limits of agreement were 2 (-10 to 14) and 1 (-4 to 6) for the left and right Apple Watches (Figure 2).

INSERT FIGURE 1 ABOUT HERE

INSERT FIGURE 2 ABOUT HERE

Inter-device standardised typical error and ICC are displayed in Figure 3. Individual variation in HR_{max} across the devices compared with age-predicted calculations for HR_{max} are presented in Figure 4.

INSERT FIGURE 3 ABOUT HERE

INSERT FIGURE 4 ABOUT HERE

Discussion

This is the first time that the validity and inter-device variability of the Apple Watch™ for measuring HR_{max} has been investigated. The Apple Watch™ displayed good to very good criterion validity (Hopkins, 2016) for measuring HR_{max}, an important component of accurate exercise prescription, compared to the widely accepted Polar T31™ heart rate monitor. The data in our study are largely in agreement with that observed by Wallen et al. (2016) who reported the 95% limits of agreement as -10 to 7 beats·min^{-1}. The limits of agreement for the left and right Apple Watch™ in our study fall either side of these values, with the left watch showing wider limits and the right watch narrower limits. Although the current study has a slightly smaller sample size it is strengthened by the direct access to the raw heart rate data and the watches being worn simultaneously during testing, which neither Wallen et al. (2016) or Wang et al. (2016) reported doing. Inter-device variability in HR_{max} measured by the Apple Watch™ is moderate when worn simultaneously on different arms. It is unclear why one individual had a larger inter-device variability (Figure 3), which warrants further investigation as this outlier appears largely responsible for the greater variability in the Apple Watch™ worn on the left wrist. Given the error that can be caused by motion artefact (Allen, 2007), it is possible that the arm movement of this individual was considerably different from other participants, although we have no objective data to confirm or refute this assertion.

Given the interest in consumer-based sensor and wearable technology (IDC, 2016) it is important to have accurate maximal heart rate measurements for exercise prescription, especially given the implications for user safety, motivation and adherence. The Apple Watch™ was within the range of the typical variability associated with using the 220 - age formula (10-12 beats·min^{-1}) and 206.9 - (0.67 x age) formula (5-8 beats·min^{-1}) (Gellish et al., 2007) and within 3 beats·min^{-1} of the mean chest strap HR_{max} demonstrating that using the Apple Watch™ is an acceptable alternative method.

Unlike measuring HR manually, both the Polar T31™ HR strap and the Apple Watch™ enable continuous and immediate feedback of HR during exercise, which has the potential to enhance self-regulation, exercise safety and motivation (Lyons et al., 2014). Additionally, it could improve the ability to adhere to the exercise prescription. This is particularly important when individuals, such as those with chronic diseases, need to stay under specific maximal heart rate thresholds recommended by a health professional for safety reasons (Price, Gordon, Bird, & Benson, 2016). To that end, mobile health technology has been reported to facilitate better management and improved patient confidence in monitoring their condition in chronic disease populations (Hamine, Gerth-Guyette, Faulx, Green, & Ginsburg, 2015).

Conclusions

The Apple Watch™ has good to very good criterion validity for measuring HR_{max}, with no substantial under- or over-estimation. There were moderate and small prediction errors for the left and right watches, respectively. Inter-device variability in HR_{max} is moderate. Users need to weigh up the validity and variability of the device compared with the associated cost of the Apple Watch™ or chest strap in determining what is most suitable for their needs.

Conflicts of interest: None

Acknowledgements: No funding was received to conduct this study.

References

Achten, J., & Jeukendrup, A. E. (2003). Heart rate monitoring: applications and limitations. *Sports Medicine, 33*(7), 517-538

Allen, J. (2007). Photoplethysmography and its application in clinical physiological measurement. *Physiological Measurement, 28*(3), R1-39. doi: 10.1088/0967-3334/28/3/R01

Canalys. (2016). *Apple shipped two-thirds of all smart watches in 2015*. Retrieved from https://www.canalys.com/static/press_release/2016/media-alert-05022016-apple-shipped-two-thirds-all-smart-watches-2015.pdf

Gellish, R. L., Goslin, B. R., Olson, R. E., Mcdonald, A., Russi, G. D., & Moudgil, V. K. (2007). Longitudinal modeling of the relationship between age and maximal heart rate. *Medicine and Science in Sports and Exercise, 39*(5), 822-829. doi: 10.1097/mss.0b013e31803349c6

Hamine, S., Gerth-Guyette, E., Faulx, D., Green, B. B., & Ginsburg, A. S. (2015). Impact of mHealth chronic disease management on treatment adherence and patient outcomes: a systematic review. *Journal of Medical Internet Research, 17*(2), e52. doi: 10.2196/jmir.3951

Hertzman, A. B. (1938). The blood supply of various skins areas as estimated by the photoelectric plethysmograph. *American Journal of Physiology, 124*, 328-340

Hopkins, W. G. (2007). Estimating sample size for magnitude-based inferences. *Sportscience, 10*, 63-68

Hopkins, W. G. (2015). Spreadsheets for analysis of validity and reliability. *Sportscience*, 36-42

Hopkins, W. G. (2016). *Validity thresholds and error rates for test measures used to assess individuals*. 21st Annual Congress of the European College of Sport Science, Vienna, Austria.

Howley, E. T., Bassett, D. R., & Welch, H. G. (1995). Criteria for maximal oxygen uptake: review and commentary. *Medicine and Science in Sports and Exercise, 27*(9), 1292-1301

IDC. (2016). *IDC Forecasts Wearables Shipments to Reach 213.6 Million Units Worldwide in 2020 with Watches and Wristbands Driving Volume While Clothing and Eyewear Gain Traction*. Retrieved from http://www.idc.com/getdoc.jsp?containerId=prUS41530816

Lyons, E. J., Lewis, Z. H., Mayrsohn, B. G., & Rowland, J. L. (2014). Behavior change techniques implemented in electronic lifestyle activity monitors: a systematic content analysis. *Journal of Medical Internet Research, 16*(8), e192. doi: 10.2196/jmir.3469

Norton, K. I., Marfell-Jones, M. J., Whittingham, M., Kerr, D., Carter, L., Saddington, K., & Gore, C. J. (2000). Anthropometric assessment protocols. In C. J. Gore (Ed.), *Physiological Tests for Elite Athletes*. Champaign, IL.: Human Kinetics.

Pescatello, L. S., & American College of Sports Medicine. (2014). *ACSM's Guidelines for Exercise Testing and Prescription* (9th ed.). Philadelphia, PA: Wolters Kluwer/Lippincott Williams & Wilkins Health.

Price, K. J., Gordon, B. A., Bird, S. R., & Benson, A. C. (2016). A review of guidelines for cardiac rehabilitation exercise programmes: Is there an international consensus? *European Journal of Preventive Cardiology, 23*(16), 1715-1733. doi: 10.1177/2047487316657669

Robergs, R. A., & Landwehr, R. (2002). The surprising history of the "HRmax=220-age" equation. *Journal of Exercise Physiology Online, 5*(2), 1-10

Wallen, M. P., Gomersall, S. R., Keating, S. E., Wisløff, U., & Coombes, J. S. (2016). Accuracy of Heart Rate Watches: Implications for Weight Management. *PloS One, 11*(5), e0154420. doi: 10.1371/journal.pone.0154420

Wang, R., Blackburn, G., Desai, M., Phelan, D., Gillinov, L., Houghtaling, P., & Gillinov, M. (2016). Accuracy of Wrist-Worn Heart Rate Monitors. *JAMA Cardiology*. doi: 10.1001/jamacardio.2016.3340

Figure legends

Figure 1. Correlation of maximal heart rate between the Polar™ heart rate strap (criterion) and left (A) and right (B) Apple Watch™.

Figure 2. Bland-Altman plots showing the mean bias and 95% limits of agreement for maximal heart rate derived from a left (A) and right (B) Apple Watch^{TM} compared to the criterion Polar T31^{TM}.

Figure 3. Inter-device (Apple Watch™ worn on the right and left wrist) standardised typical error (TE) (A) and intraclass correlation (ICC) (B).

Figure 4. Individual variation in HR_{max} across the devices (Apple Watch™ worn on left and right wrist, Polar™ heart rate strap,) compared with age-predicted calculations. Each black bar represents the group mean.