PLOS One

RESEARCH ARTICLE

Investigating the accuracy of Apple Watch VO_2 max measurements: A validation study

Rory Lambe[1,2*], **Ben O'Grady**[1,2], **Maximus Baldwin**[1,2,3], **Cailbhe Doherty**[1,2]

1 School of Public Health, Physiotherapy and Sports Science, University College Dublin, Dublin 4, Ireland, **2** Insight Research Ireland Centre for Data Analytics, University College Dublin, Dublin 4, Ireland, **3** Institute for Sport and Health, University College Dublin, Dublin 4, Ireland

* rory.lambe@ucdconnect.ie

OPEN ACCESS

Citation: Lambe R, O'Grady B, Baldwin M, Doherty C (2025) Investigating the accuracy of Apple Watch VO_2 max measurements: A validation study. PLoS One 20(5): e0323741. https://doi.org/10.1371/journal.pone.0323741

Editor: Emiliano Cè, Università degli Studi di Milano: Universita degli Studi di Milano, ITALY

Received: January 29, 2025

Accepted: April 14, 2025

Published: May 15, 2025

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pone.0323741

Copyright: © 2025 Lambe et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract

VO_2 max is a measure of cardiorespiratory fitness and a key indicator of overall health. It is predictive of cardiovascular events and shows a strong inverse association with all-cause mortality. Increased cardiorespiratory fitness is associated with reductions in coronary artery disease, diabetes and cancer. Apple Watch offers a less resource-intensive and more feasible alternative to the gold standard assessment for VO_2 max, indirect calorimetry, but the accuracy of its measurements remains uncertain. This study aimed to assess the validity of VO_2 max estimates from Apple Watch in comparison to indirect calorimetry. Thirty participants wore an Apple Watch for 5-10 days to generate a VO_2 max estimate. Subsequently, they underwent a maximal exercise treadmill test in accordance with the modified Åstrand protocol. The agreement between measurements from Apple Watch and indirect calorimetry was assessed using Bland-Altman analysis, mean absolute percentage error (MAPE), and mean absolute error (MAE). Overall, Apple Watch underestimated VO_2 max, with a mean difference of 6.07 mL/kg/min (95% CI 3.77–8.38). Limits of agreement indicated variability between measurement methods (lower -6.11 mL/kg/min; upper 18.26 mL/kg/min). MAPE was calculated as 13.31% (95% CI 10.01–16.61), and MAE was 6.92 mL/kg/min (95% CI 4.89–8.94). These findings indicate that Apple Watch VO_2 max estimates require further refinement prior to clinical implementation. However, further consideration of Apple Watch as an alternative to conventional VO_2 max prediction from submaximal exercise is warranted, given its practical utility.

Introduction

VO_2 max, introduced by British physiologists Hill and Lupton, is the maximum amount of oxygen an individual can utilise during exercise [1]. It provides an evaluation of cardiorespiratory fitness and has been established as an important barometer of health [2–4]. Epidemiological evidence supports an inverse and independent association between cardiorespiratory fitness and mortality, as measured by VO_2 max or

Data availability statement: All data files are available from the primary author's GitHub repository at github.com/rorylambe/applewatch-validation

Funding: This project was funded by Science Foundation Ireland's National Challenge Fund (grant ID: 22/NCF/FD/10949). The funding was received by CD. The funder had no role in any aspect of the trial. Funder URL: www.sfi.ie

Competing interests: The authors have declared that no competing interests exist.

metabolic equivalent (MET) [5–7]. It is a predictor of cardiovascular events [8–10], and some studies report it to be more predictive of cardiovascular and all-cause mortality than well-known risk factors such as hypertension, obesity, and hypercholesterolemia [11–13]. In recognition of its significance, the American Heart Association advocates for the routine assessment of cardiorespiratory fitness and has proposed it as a clinical vital sign [14].

VO_2 max is the most widely accepted measure of cardiorespiratory fitness, and the gold standard for its assessment is indirect calorimetry [15]. Indirect calorimetry measures maximal oxygen uptake by recording oxygen consumption and carbon dioxide production during cardiopulmonary exercise testing (CPET) [16]. While it is the most accurate form of measurement, it is costly and resource-intensive, limiting its use in clinical practice and among the general population [17]. To overcome these difficulties, VO_2 max is often derived from submaximal exercise tests that require less time and are considered lower-risk [18]. Error for VO_2 max predictions from submaximal tests is reported to range between 1.6 and 4.1 mL/kg/min [19,20].

Wearable devices also use submaximal exercise to derive VO_2 max [17], and their recent proliferation has democratised monitoring of physiological measurements. They provide unobtrusive, longitudinal monitoring outside the confines of exercise physiology laboratories or healthcare settings, and may facilitate remote monitoring of cardiorespiratory fitness [21,22]. The widespread uptake of wearable devices has been catalysed by the ready availability of a suite of health-related metrics, and their increasing prominence is illustrated by the American College of Sports Medicine (ACSM), who identified wearable technology as the number one fitness trend for 2025 [23]. Over half the population in many countries now owns a wearable device – ranging from smartwatches to smart rings – with global user numbers expected to grow significantly in the coming years [24]. The wearable technology market is projected to reach a valuation of $186 billion by 2030 [25], and Apple Watch holds the largest market share with over 100 million users worldwide [26].

The first-generation Apple Watch was released in 2015, and its VO_2 max estimation feature was introduced with watchOS 4 in September 2017 [27]. This enabled Apple Watch Series 3 and later to estimate VO_2 max using data such as heart rate, Global Navigation Satellite System-derived metrics (e.g., speed, distance, and elevation), and user demographics such as sex, age, height, and weight [28]. Apple subsequently updated its VO_2 max algorithm in 2021 as part of watchOS 7 [28]. While this feature evoked the prospect of an assessment method that could be conducted independently of healthcare or exercise professionals, it was accompanied by notable challenges.

The accuracy of wearable-derived VO_2 max estimates represents one of the most significant challenges. A multitude of factors can influence measurements: physiological factors such as heart rate response to exercise and perfusion; environmental factors like skin contact and motion [29]; or external influences, including caffeine intake and carrying additional weight beyond body weight, such as bags or equipment [28]. Compounding these issues is the proprietary and evolving nature of the algorithms used by wearable devices. With each new hardware or software update, algorithms are subject to change, necessitating recurrent validation by researchers [18,30,31]. This validation is hindered by the protracted process of conducting and publishing research. Consequently, fewer than 5% of consumer wearables have been validated for the full range of biometric outcomes they measure, including VO_2 max [18]. In fact, just one study evaluating the accuracy of Apple Watch for VO_2 max estimation has been published to date [32].

Evaluating the accuracy of VO_2 max estimates is fundamental to integrating wearable devices into individual health monitoring and research contexts [30]. For consumers, the reliability and validity of these measurements can inform decisions about personal health and fitness. For researchers, wearables hold the potential to revolutionise public health surveillance by enabling the collection of large-scale, population-level data without the logistical and financial constraints of laboratory-based methods [33]. On the contrary, inaccurate or inconsistent data may compromise the validity of such research, particularly when used to inform health policy or population health interventions. Validation is required to seal the fissure between technological innovation and evidence-based application in health and fitness settings.

The aim of this study was to evaluate the accuracy of VO_2 max estimates generated by Apple Watch in comparison to the gold standard method, indirect calorimetry. Specifically, the objectives were to assess the level of agreement between the two methods, and to quantify the degree of error in VO_2 max estimates from Apple Watch.

Materials and Methods

Study design and oversight

This cross-sectional validation study was conducted in Dublin, Ireland between December 2023 and July 2024. Recruitment commenced on Monday, 4 December 2023, and finished on Monday, 22 July 2024. Healthy adults, aged 18 or older, were recruited through social media, posters, and word of mouth. Individuals with cardiovascular or mental illness, and those using medication affecting cardiovascular function were not eligible for inclusion. All participants completed the Physical Activity Readiness Questionnaire Plus (PAR-Q+) to ensure suitability for exercise, and written informed consent was obtained. Ethical approval was granted by the University College Dublin Human Research Ethics Committee (reference number: LS-23-55) on November 15th, 2023.

Generating VO_2 max estimates with Apple Watch

Apple Watch generates VO_2 max estimates (proprietarily referred to as ‘Cardio Fitness’) by measuring an individual’s heart rate response to exercise during outdoor walking, running, or hiking activities on ground of less than 5% incline or decline [28]. Adequate GPS and heart rate signal must be obtained, alongside an increase of approximately 30% in heart rate from the resting value [28]. Multiple activities are required to generate an estimate, although this number varies between users [28]. Participants were informed of the required procedure and were requested to generate an estimate independently within one week of criterion testing. The process of generating a VO_2 max estimate was conducted in accordance with manufacturer guidelines [28]. Participants were instructed to accurately input all required demographic information in the Health app prior to completing any exercise activities, including height, weight, sex, and age. If participants did not already own an Apple Watch, they were provided with an Apple Watch Series 9 or Ultra 2 for a period of 5-10 days. All Apple Watch devices were updated to watchOS 10 or later, ensuring that this validation study exclusively assessed devices using Apple’s latest VO_2 max prediction algorithm and the most recent software version.

Indirect calorimetry testing

Subsequently, each participant underwent a maximal cardiopulmonary exercise test (CPET) using indirect calorimetry at the Institute for Sport and Health, University College Dublin. Indirect calorimetry is considered the gold standard for VO_2max testing [1,17]. An exercise treadmill test was conducted in accordance with the modified Åstrand protocol [34]. A treadmill speed between 8 and 13 km/h was selected and this remained consistent throughout [34]. The incline of the treadmill was increased by 2.5% every two minutes following the initial three-minute warm-up period at 0% incline [34]. The COSMED Quark CPET metabolic cart (COSMED, Trentino, Italy) was calibrated prior to each test according to manufacturer instructions and testing was conducted on the h/p/cosmos Venus treadmill (h/p/cosmos, Nußdorf-Traunstein, Germany). To ensure true VO_2 max had been attained, participants were required to meet at least two of the following criteria: heart rate within ±10 bpm of age-predicted maximum (220 – age); respiratory exchange ratio (RER) ≥ 1.15; rate of perceived exertion (RPE) ≥ 17; VO_2 plateau, defined as an increase in VO_2 of less than 150 mL/min/kg, with an increase in work rate as evidenced by physiological data [35]. If at least two of these criteria were not met, the value was regarded as a VO_2 peak. Participants were instructed to refrain from caffeine and nicotine for 12 hours prior to testing, and to avoid strenuous activity and alcohol for a minimum of 24 hours [15].

Outcomes

The primary outcome was the agreement between VO_2 max estimates derived from Apple Watch with those obtained via indirect calorimetry, which was calculated using Bland-Altman limits of agreement analysis [36], mean absolute percentage error (MAPE), and mean absolute error (MAE).

Statistical analysis

The target sample size was calculated based on the sample sizes of previous studies investigating the measurement accuracy of other consumer wearable devices [32,37,38]. Additional participants were not recruited to account for potential dropouts due to the cross-sectional nature of the study design. Data from the COSMED Quark CPET were filtered using a time average of 30 seconds and exported to Microsoft Excel. The highest time-averaged VO_2/kg value was interpreted as VO_2 max. This was compared to the most recent Apple Watch VO_2 max estimate available in the Health app. Participants who did not attain VO_2 max were not included in the final analysis.

To evaluate the agreement between Apple Watch and the criterion, Bland-Altman limits of agreement analysis was conducted. The mean difference (bias) and 95% limits of agreement (LoA), defined as the mean difference ±1.96 times the standard deviation of the differences, were calculated. A Bland-Altman plot was generated to visually evaluate the agreement. Mean absolute percentage error (MAPE) and mean absolute error (MAE) were also calculated. The t-critical values were used to compute corresponding 95% confidence to account for the sample size. Analyses were conducted in Python (version 3.12) using pandas, Matplotlib, and NumPy packages. The scripts used are available at github.com/rorylambe/applewatch-validation.

Results

Baseline characteristics

A total of 30 individuals participated in this study (mean age [SD], 31.86 [13.99] years; 50% female). Of the 30 participants, two (one male, one female) did not achieve the required criteria for VO_2 max, and as such 28 participants were included in the analyses. One of the excluded participants met only the threshold for rate of perceived exertion whereas the other failed to attain any of the criteria. Both excluded participants were wearing Apple Watch Series 9. The cardiorespiratory fitness level of each participant, determined by indirect calorimetry testing, was classified in accordance with the reference standard from the Fitness Registry and the Importance of Exercise National Database (FRIEND) [39]. 21 participants were classified as having either 'Superior' or 'Excellent' cardiorespiratory fitness, while nine individuals were classified as 'Good' or 'Fair'. Participant baseline characteristics, including BMI and Fitzpatrick skin tone, are listed in Table 1.

Apple Watch agreement with the criterion

Overall, Apple Watch underestimated VO_2 max in comparison to indirect calorimetry. The mean difference was 6.07 mL/kg/min (SD 6.22; 95% confidence interval [CI] 3.77–8.38). Bland-Altman limits of agreement (LoA) indicated variability

Table 1. Characteristics of the participants at baseline.

Characteristic	Value
Age, mean (SD), years	31.86 (13.99)
Female, no. (%)	15 (50%)
Height, mean (SD), cm	172.23 (7.85)
Weight, mean (SD), kg	70.55 (9.13)
BMI, mean (SD), kg/m^2	23.76 (2.54)
Fitzpatrick skin tone [a]	
Type I	4
Type II	22
Type III	4
CRF level [b]	
Superior, no.	3
Excellent, no.	18
Good, no.	8
Fair, no.	1
Apple Watch model [c]	
Series 9	17
Ultra 2	6
Series 8 and SE 2	2 of each
Series 5, 6, 7	1 of each

BMI: Body Mass Index

[a] Number of participants of each skin tone according to the Fitzpatrick Scale.

[b] Number of participants of each cardiorespiratory fitness level classified according to FRIEND.

[c] Number of participants who wore each Apple Watch model during the study.

https://doi.org/10.1371/journal.pone.0323741.t001

between the two measurement methods (lower LoA -6.11 mL/kg/min; upper LoA 18.26 mL/kg/min). The Bland-Altman plot is presented in Fig 1. The mean absolute percentage error (MAPE) was 13.31% (95% CI 10.01–16.61); mean absolute error (MAE) was 6.92 mL/kg/min (95% CI 4.89–8.94). To assess statistical power and effect size, a one-sample t-test was conducted comparing Apple Watch and criterion VO_2 max values. This analysis yielded a t-value of –5.07 (df = 27, $p < 0.001$), indicating a significant difference with a large effect size (Cohen’s d = –0.96). A post-hoc power analysis (α = 0.05) confirmed the sample size (n = 28) provided >99% power, suggesting the study was sufficiently powered to detect a true difference between the measurement methods, thereby minimising the risk of Type II error. No trends in accuracy could be observed based on Apple Watch model. Results are summarised in Table 2.

Discussion

This validation study evaluated the accuracy of VO_2 max estimates from Apple Watch compared to the gold-standard method of indirect calorimetry. Overall, Apple Watch underestimated VO_2 max, with a mean difference of 6.07 mL/kg/min (SD 6.22) and a MAPE of 13.31%. Bland-Altman limits of agreement indicated variability between the two measurement methods (lower -6.11 mL/kg/min; upper 18.26 mL/kg/min).

These findings align with those of a prior study evaluating Apple Watch Series 7, which reported a similar mean underestimation of 4.5 mL/kg/min [32]. However, important methodological differences exist. In the previous study, criterion VO_2 max values were obtained through a graded exercise test conducted on a cycle ergometer, rather than using a treadmill-based test. Apple Watch derives VO_2 max from walking, running or hiking activities [28], and comparison of VO_2

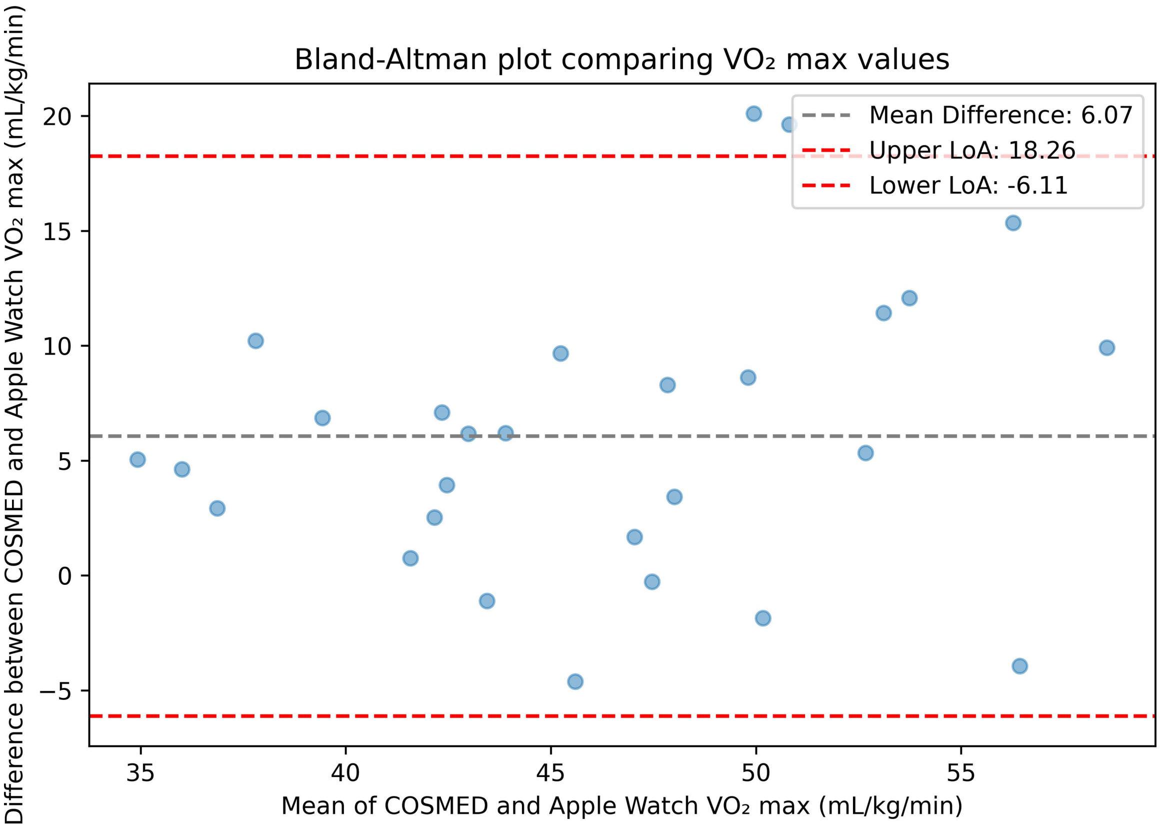

Fig 1. Bland-Altman plot demonstrating the agreement between VO2 max measurements from Apple Watch and the criterion measure. Upper LoA = upper limit of agreement; Lower LoA = lower limit of agreement.

https://doi.org/10.1371/journal.pone.0323741.g001

Table 2. Statistical measures of agreement between Apple Watch and indirect calorimetry.

Statistical measure	Result
Mean (SD), mL/kg/min	
Apple Watch	43.27 (6.34)
COSMED*	49.35 (7.79)
Standard error of the mean, mL/kg/min	
Apple Watch	1.20
COSMED	1.47
Standard deviation of the differences, mL/kg/min	6.22
Mean difference (95% CI), mL/kg/min	6.07 (3.77–8.38)
Bland-Altman limits of agreement (lower, upper), mL/kg/min	−6.11, 18.26
Mean absolute percentage error (95% CI)	13.31% (10.01–16.61)
Mean absolute error, mL/kg/min (95% CI)	6.92 (4.89–8.94)

* COSMED = COSMED Quark CPET metabolic cart.

https://doi.org/10.1371/journal.pone.0323741.t002

max measured through different exercise modalities – such as cycling compared with running – is not advised due to the inherent discrepancies in results they produce [40]. Typically, treadmill tests yield higher VO_2 max values, although the difference is smaller in well-trained cyclists [41]. This difference underpins the INTERLIVE consortium's recommendation that criterion testing methods should closely align with the activity used by the wearable device to generate VO_2 max predictions [17].

While these methodological differences are important to consider, the authors noted distinct trends based on the fitness levels of participants [32]. For individuals with good or excellent fitness, Apple Watch demonstrated a propensity to underestimate VO_2 max, whereas among those with poor fitness, there was a tendency to overestimate. Although these authors found an overall underestimation, a separate meta-analysis of fourteen studies, which included multiple wearable device brands, reported a minor overestimation (pooled bias -0.09 mL/kg/min) [17]. Despite a very low bias, the limits of agreement were comparatively wide (lower -9.92 mL/kg/min; upper 9.74 mL/kg/min), suggesting both positive and negative differences in estimates of a relatively large magnitude, which collectively yielded a bias close to zero. A propensity to both under- and overestimate may arise from prediction algorithms that are developed using population-level data or mean values for specific demographics. Given our study sample predominantly consisted of individuals with high cardiorespiratory fitness levels, this may constitute one explanation for our finding of underestimation. Importantly, it highlights the need to consider participant demographic characteristics, as well as multiple statistical measures of agreement, when interpreting results.

Our findings, combined with the broader literature, underscore the challenge of deriving maximal oxygen consumption from submaximal exercise for wearable devices. Like conventional prediction equations, Apple's prediction algorithm is founded on heart rate response to exercise [28]. The inclusion of machine learning techniques and additional motion sensor measurements is also probable, although the precise nature of the algorithm remains undisclosed [30]. In clinical settings, submaximal exercise testing is often used alongside prediction equations to estimate VO_2 max, owing to safety and feasibility concerns [35]. These equations predict VO_2 max based on an individual's physiological response to exercise of a quantified workload, and six assumptions – outlined by the ACSM – must be achieved to ensure optimal accuracy [35]. They include a minimal difference between actual and predicted maximal heart rate, knowledge of medication or substances affecting heart rate, and the existence of a linear relationship between heart rate and work rate. Yet, accuracy from traditional prediction equations can vary significantly [42–44]. A study (n = 541) evaluating three different submaximal exercise tests for estimating maximal oxygen uptake reported a standard error of the estimation ranging from 3.7 to 4.5 mL/kg/min [45]. In a separate study examining the Chester Step Test, the predicted aerobic capacity had a standard error of 3.9 mL/kg/min [19,46]. Comparatively, measurement error for laboratory-based indirect calorimetry has been estimated at ±5%, with a meta-analysis of 39 studies reporting a mean standard error of 2.58 mL/kg/min [40,47]. Considering this literature, the results of our validation study indicate that Apple Watch VO_2 max predictions fall approximately two to three standard deviations beyond the typical error of the gold standard measurement. However, they align more closely with the magnitude of error reported for traditional VO_2 max predictions derived from submaximal exercise.

While machine learning and continuous monitoring are distinct merits of Apple Watch compared to traditional prediction methods, the uncontrolled nature of the exercise used to generate Apple Watch VO_2 max estimates presents a considerable challenge. Limited information of external factors that affect heart rate response to exercise – including exercise surface, environmental temperature, or alcohol intake – can adversely impact predictions [28]. Moreover, heart rate response to lower-intensity exercise – often used to estimate VO_2 max – varies considerably between individuals [48]. Estimates based on such exercise may misinform predictions. Due to the proprietary nature of Apple's algorithm, we remain unaware of the specific data, or features, used to train and develop the prediction model. Therefore, we cannot be certain what factors most influence accuracy, or in which instances predictions are most reliable [30]. Considering this, a defined testing protocol – requiring users to input specific information and conduct a particular type of exercise – may reduce the level of uncertainty and error. Despite the imprecision of predictions from submaximal exercise – both from traditional testing methods and wearable devices – its widespread clinical use illustrates its value. The potential for Apple Watch to provide a commensurate assessment method that can be conducted independently by members of the general public warrants careful consideration. To investigate this, validation of contemporary technology is apt.

Continual, or ‘living’, validation is essential to keep pace with the iterative commercial ecosystem. The annual update cycle of Apple Watch poses a significant challenge to evaluating current technology, and the prolonged academic publication process often means that the hardware or software under validation has been discontinued by the time of publication [18,30]. This is accentuated by the rapid advancement of machine learning – and the associated updates to watchOS – which play an increasingly important role in providing accurate measurements due to the complexity of interpreting multiple sensor measurements and sensitive photoplethysmography waveforms [49,50]. Amidst these challenges, this study provides a timely evaluation of Apple’s most recent optical heart rate sensor and its latest VO_2 max prediction algorithm [28,51]. It is the first study to evaluate Apple Watch Series 9 and Ultra 2. As wearable technology evolves, agile validation will be critical to realising the potential of consumer wearables in large-scale public health applications and personalised fitness monitoring.

Central to this potential is the self-directed nature of wearable device measurements, and this study’s methodology was designed to reflect this. Participants were not required to adhere to a prescribed protocol to generate a VO_2 max prediction. Rather, they were solely informed of the requirements, meaning that estimates were generated through activities that more accurately reflected each individual’s real-world use, enhancing the ecological validity of our findings. Additionally, a robust statistical approach was used, guided by recommendations of the INTERLIVE expert consortium [17], and a sex-balanced study sample was recruited. However, the predominance of participants with high levels of cardiorespiratory fitness represents the study’s most significant limitation. Only one participant was classified as having ‘Fair’ cardiorespiratory fitness, while all others were categorised as ‘Good’, ‘Excellent’, or ‘Superior’. This limits the generalisability of our findings to populations with lower fitness levels, including clinical cohorts. Additionally, the sample primarily consisted of younger individuals, despite efforts to recruit participants across a broad age range. Another limitation is the uncontrolled procedure of generating Apple Watch VO_2 max estimates. While this protocol reflects real-world use, it introduces unquantified external variables that may have influenced prediction accuracy. Lastly, only one VO_2 max estimate was collected per participant, preventing analysis of intra-subject variability, which may have provided insights into reliability and changes in measurement accuracy over time.

The findings of this study have important implications for the use of wearable devices such as Apple Watch in clinical and research contexts. VO_2 max estimates from Apple Watch offer a cost-effective and scalable alternative to laboratory-based testing, enabling studies involving diverse populations, as well as digital endpoints for clinical trials. However, given the observed variability and underestimation of VO_2 max in this study, caution is warranted when interpreting the data in clinical or research settings. While Apple Watch may provide a suitable estimation of aerobic capacity for general fitness monitoring, our findings suggest that estimates are not sufficiently accurate to inform clinical decision-making. Future research should include individuals with diverse cardiorespiratory fitness levels to enhance generalisability, support the refinement of predictive algorithms, and develop frameworks for ongoing evaluation of wearable devices. Such efforts are essential to bridging the gap between technological innovation and evidence-based practice, ultimately enhancing the utility of wearable devices for individual health monitoring and public health applications.

Conclusions

This study found that Apple Watch underestimated VO_2 max compared to indirect calorimetry in a sample predominantly composed of individuals with high cardiorespiratory fitness levels. The margin of error more closely aligned with that of VO_2 max derived from conventional submaximal exercise testing than with VO_2 max obtained via indirect calorimetry. The unstructured nature of the exercise used to generate Apple Watch VO_2 max predictions introduces external factors – such as environmental conditions and variations in heart rate response to exercise – that are not fully accounted for by the device. Nevertheless, Apple Watch holds promise as a practical and accessible alternative to conventional submaximal exercise testing. Its potential clinical utility warrants further investigation, and ongoing validation is required to evaluate its accuracy as hardware and software evolve. Future research should focus on expanding validation efforts to include diverse populations, ensuring these devices can be effectively integrated into both individual health monitoring and broader public health applications.

Author contributions

Conceptualization: Rory Lambe, Ben O'Grady, Maximus Baldwin, Cailbhe Doherty.

Data curation: Rory Lambe, Ben O'Grady, Maximus Baldwin, Cailbhe Doherty.

Formal analysis: Rory Lambe.

Funding acquisition: Cailbhe Doherty.

Investigation: Rory Lambe, Ben O'Grady, Cailbhe Doherty.

Methodology: Rory Lambe, Ben O'Grady, Maximus Baldwin, Cailbhe Doherty.

Project administration: Rory Lambe, Ben O'Grady, Maximus Baldwin, Cailbhe Doherty.

Software: Rory Lambe.

Supervision: Cailbhe Doherty.

Validation: Rory Lambe, Ben O'Grady, Maximus Baldwin.

Visualization: Rory Lambe.

Writing – original draft: Rory Lambe, Ben O'Grady, Maximus Baldwin, Cailbhe Doherty.

Writing – review & editing: Rory Lambe, Maximus Baldwin, Cailbhe Doherty.

References

1. Hill A, Lupton H. Muscular exercise, lactic acid, and the supply and utilization of oxygen. Quar J. Med. 1923; (62):135-71.
2. Kokkinos P, Faselis C, Samuel IBH, Pittaras A, Doumas M, Murphy R, et al. Cardiorespiratory fitness and mortality risk across the spectra of age, race, and sex. J. Am. Coll. Cardiol. 2022;80(6):598-609.
3. Carnethon MR, Gidding SS, Nehgme R, Sidney S, Jacobs Jr DR, Liu K. Cardiorespiratory fitness in young adulthood and the development of cardiovascular disease risk factors. Jama. 2003;290(23):3092-100.
4. Lee D-c, Artero EG, Sui X, Blair SN. Mortality trends in the general population: the importance of cardiorespiratory fitness. Journal of psychopharmacology. 2010;24(4_suppl):27-35.
5. Sanchis-Gomar F, Lavie CJ, Marín J, Perez-Quilis C, Eijsvogels TM, O'Keefe JH, et al. Exercise effects on cardiovascular disease: from basic aspects to clinical evidence. Cardiovascular Research. 2022;118(10):2253-66.
6. Myers J, Prakash M, Froelicher V, Do D, Partington S, Atwood JE. Exercise capacity and mortality among men referred for exercise testing. New England journal of medicine. 2002;346(11):793-801.
7. Kokkinos PF, Faselis C, Myers J, Panagiotakos D, Doumas M. Interactive effects of fitness and statin treatment on mortality risk in veterans with dyslipidaemia: a cohort study. The Lancet. 2013;381(9864):394-9.
8. Blair SN, Kohl HW, Paffenbarger RS, Clark DG, Cooper KH, Gibbons LW. Physical fitness and all-cause mortality: a prospective study of healthy men and women. JAMA 1989;262(17):2395-401.
9. Mandsager K, Harb S, Cremer P, Phelan D, Nissen SE, Jaber W. Association of cardiorespiratory fitness with long-term mortality among adults undergoing exercise treadmill testing. JAMA network open. 2018;1(6):e183605-e.
10. Clausen JS, Marott JL, Holtermann A, Gyntelberg F, Jensen MT. Midlife cardiorespiratory fitness and the long-term risk of mortality: 46 years of follow-up. J. Am. Coll. Cardiol. 2018;72(9):987-95.
11. Mora S, Redberg RF, Cui Y, Whiteman MK, Flaws JA, Sharrett AR, et al. Ability of exercise testing to predict cardiovascular and all-cause death in asymptomatic women: a 20-year follow-up of the lipid research clinics prevalence study. Jama. 2003;290(12):1600-7.
12. Laukkanen JA, Kurl S, Salonen R, Rauramaa R, Salonen JT. The predictive value of cardiorespiratory fitness for cardiovascular events in men with various risk profiles: a prospective population-based cohort study. European heart journal. 2004;25(16):1428-37.
13. Myers J, Nead KT, Chang P, Abella J, Kokkinos P, Leeper NJ. Improved reclassification of mortality risk by assessment of physical activity in patients referred for exercise testing. Am. J. Med. 2015;128(4):396-402.

14. Ross R, Blair SN, Arena R, Church TS, Després J-P, Franklin BA, et al. Importance of assessing cardiorespiratory fitness in clinical practice: a case for fitness as a clinical vital sign: a scientific statement from the American Heart Association. Circulation. 2016;134(24):e653-e99.

15. Bayles MP. ACSM's exercise testing and prescription. Lippincott Williams & Wilkins; 2023.

16. Society AT, Physicians ACoC. American Thoracic Society/American College of Chest Physicians statement on cardiopulmonary exercise testing. Am J Respir Crit Care Med. 2003;167:211-77.

17. Molina-Garcia P, Notbohm HL, Schumann M, Argent R, Hetherington-Rauth M, Stang J, et al. Validity of Estimating the Maximal Oxygen Consumption by Consumer Wearables: A Systematic Review with Meta-analysis and Expert Statement of the INTERLIVE Network. Sports Med. 2022;52(7):1577-97. https://doi.org/10.1007/s40279-021-01639-y PMID: 35072942.

18. Doherty C, Baldwin M, Keogh A, Caulfield B, Argent R. Keeping Pace with Wearables: A Living Umbrella Review of Systematic Reviews Evaluating the Accuracy of Consumer Wearable Technologies in Health Measurement. Sports Medicine. 2024:1-20.

19. Sykes K, Roberts A. The Chester step test—a simple yet effective tool for the prediction of aerobic capacity. Physiotherapy. 2004;90(4):183-8. https://doi.org/10.1016/j.physio.2004.03.008

20. Buckley J, Sim J, Eston R, Hession R, Fox R. Reliability and validity of measures taken during the Chester step test to predict aerobic power and to prescribe aerobic exercise. Br. J. Sports Med. 2004;38(2):197-205.

21. Jia Y, Wang W, Wen D, Liang L, Gao L, Lei J. Perceived user preferences and usability evaluation of mainstream wearable devices for health monitoring. PeerJ. 2018;6:e5350.

22. Hickey AM, Freedson PS. Utility of consumer physical activity trackers as an intervention tool in cardiovascular disease prevention and treatment. Prog. Cardiovasc. Dis. 2016;58(6):613-9.

23. A'Naja MN, Batrakoulis A, Camhi SM, McAvoy C, Sansone JS, Reed R. 2025 ACSM worldwide fitness trends: future directions of the health and fitness industry. ACSM's Health Fit. J. 2024;28(6):11-25.

24. Wearable devices ownership in selected countries as of March 2024 [Internet]. 2024 [cited 10/04/2025]. Available from: https://www.statista.com/forecasts/1101101/wearable-devices-ownership-in-selected-countries.

25. Wearable Technology Market Size, Share & Trends Analysis Report By Product (Head & Eyewear, Wristwear), By Application (Consumer Electronics, Healthcare), By Region (Asia Pacific, Europe), And Segment Forecasts, 2023 - 2030. Available from: https://www.grandviewresearch.com/industry-analysis/wearable-technology-market: Grand View Research, 2023.

26. Apple Watch Unit Sales Worldwide 2015-2023 [Internet]. Statista. 2024 [cited 25/09/2024]. Available from: https://www.statista.com/statistics/1421546/apple-watch-sales-worldwide/.

27. About watchOS 4 Updates Apple Support: Apple; 2017 [cited 2024 13/11/2024]. Available from: https://support.apple.com/en-us/111739.

28. Using Apple Watch to Estimate Cardio Fitness with VO2 max. Apple Healthcare: Apple, 2021.

29. Bent B, Goldstein BA, Kibbe WA, Dunn JP. Investigating sources of inaccuracy in wearable optical heart rate sensors. NPJ Digital Medicine. 2020;3(1):18.

30. Schumann M, Doherty C. Bridging Gaps in Wearable Technology for Exercise and Health Professionals: A Brief Review. Int J Sports Med. 2024. https://doi.org/10.1055/a-2376-6332 PMID: 39079705.

31. O'Grady B, Lambe R, Baldwin M, Acheson T, Doherty C. The Validity of Apple Watch Series 9 and Ultra 2 for Serial Measurements of Heart Rate Variability and Resting Heart Rate. Sensors. 2024;24(19):6220.

32. Caserman P, Yum S, Göbel S, Reif A, Matura S. Assessing the Accuracy of Smartwatch-Based Estimation of Maximum Oxygen Uptake Using the Apple Watch Series 7: Validation Study. JMIR Biomed. Eng. 2024;9:e59459.

33. Keogh A, Argent R, Doherty C, Duignan C, Fennelly O, Purcell C, et al. Breaking down the Digital Fortress: The Unseen Challenges in Healthcare Technology-Lessons Learned from 10 Years of Research. Sensors. 2024;24(12). https://doi.org/10.3390/s24123780 PMID: 38931564.

34. Pollock ML, Bohannon RL, Cooper KH, Ayres JJ, Ward A, White SR, et al. A comparative analysis of four protocols for maximal treadmill stress testing. Am. Heart J. 1976;92(1):39-46.

35. Liguori G, Medicine ACoS. ACSM's Guidelines for Exercise Testing and Prescription: Lippincott Williams & Wilkins; 2020.

36. Bland JM, Altman D. Statistical methods for assessing agreement between two methods of clinical measurement. The Lancet. 1986;327(8476):307-10.

37. Freeberg KA, Baughman BR, Vickey T, Sullivan JA, Sawyer BJ. Assessing the ability of the Fitbit Charge 2 to accurately predict VO(2max). Mhealth. 2019;5:39. https://doi.org/10.21037/mhealth.2019.09.07 PMID: 31620466.

38. Passler S, Bohrer J, Blöchinger L, Senner V. Validity of Wrist-Worn Activity Trackers for Estimating VO(2max) and Energy Expenditure. Int J Environ Res Public Health. 2019;16:(17). https://doi.org/10.3390/ijerph16173037 PMID: 31443347.

39. Kaminsky LA, Arena R, Myers J, Peterman JE, Bonikowske AR, Harber MP, et al., editors. Updated reference standards for cardiorespiratory fitness measured with cardiopulmonary exercise testing: data from the Fitness Registry and the Importance of Exercise National Database (FRIEND). Mayo clinic proceedings; 2022: Elsevier.

40. Beltz NM, Gibson AL, Janot JM, Kravitz L, Mermier CM, Dalleck LC. Graded Exercise Testing Protocols for the Determination of VO(2) max: Historical Perspectives, Progress, and Future Considerations. J Sports Med (Hindawi Publ Corp). 2016;2016:3968393. https://doi.org/10.1155/2016/3968393 PMID: 28116349.

41. Millet GP, Vleck VE, Bentley DJ. Physiological differences between cycling and running: lessons from triathletes. Sports Med. 2009;39(3):179-206. https://doi.org/10.2165/00007256-200939030-00002 PMID: 19290675.

42. Evans HJ, Ferrar KE, Smith AE, Parfitt G, Eston RG. A systematic review of methods to predict maximal oxygen uptake from submaximal, open circuit spirometry in healthy adults. J Sci Med Sport. 2015;18(2):183-8. https://doi.org/10.1016/j.jsams.2014.03.006 PMID: 24721146.

43. Koutlianos N, Dimitros E, Metaxas T, Cansiz M, Deligiannis A, Kouidi E. Indirect estimation of VO2max in athletes by ACSM's equation: valid or not? Hippokratia. 2013;17(2):136-40. PMID: 24376318.

44. Kokkinos P, Kaminsky LA, Arena R, Zhang J, Franklin B, Kraus W, et al. New Equations for Predicting Maximum Oxygen Uptake in Patients With Heart Failure. Am J Cardiol. 2020;128:7-11. Epub 20200519. doi: https://doi.org/10.1016/j.amjcard.2020.04.049 PMID: 32650927.

45. Chung J, Lee K. A comparison of the validity of three exercise tests for estimating maximal oxygen uptake in Korean adults aged 19–64 years. Appl. Sci. 2022;12(3):1371.

46. Bennett H, Parfitt G, Davison K, Eston R. Validity of Submaximal Step Tests to Estimate Maximal Oxygen Uptake in Healthy Adults. Sports Med. 2016;46(5):737-50.

47. Vickers R. Measurement Error in Maximal Oxygen Uptake Tests. San Diego Report: Naval Health Research Center; 2003.

48. Bouchard C, Rankinen T. Individual differences in response to regular physical activity. Med Sci Sports Exerc. 2001;33(6 Suppl):S446-51; discussion S52-3.

49. Pollreisz D, TaheriNejad N. Detection and removal of motion artifacts in PPG signals. Mob Netw Appl 2022;27(2):728-38.

50. McHugh-Johnson M. How Google Pixel Watch 2 and Fitbit Charge 6 improved heart rate tracking blog.google2023 [28/02/2025]. Available from: https://blog.google/products/pixel/pixel-watch-2-fitbit-charge-6-heart-rate/.

51. Compare Apple Watch Models apple.com: Apple; 2024 [cited 2024 13/11/2024]. Available from: https://www.apple.com/watch/compare/.