UC San Diego
UC San Diego Previously Published Works

Title
Validity of Cardiorespiratory Fitness Measured with Fitbit Compared to V˙O2max.

Permalink
https://escholarship.org/uc/item/6cw9j8q9

Journal
Medicine and Science in Sports and Exercise, 51(11)

Authors
Klepin, Katharine
Wing, David
Higgins, Michael
et al.

Publication Date
2019-11-01

DOI
10.1249/MSS.0000000000002041

Peer reviewed

Validity of Cardiorespiratory Fitness Measured with Fitbit Compared to $\dot{V}O_{2max}$

KATHARINE KLEPIN[1], DAVID WING[2], MICHAEL HIGGINS[2], JEANNE NICHOLS[1,2], and JOB G. GODINO[2,3]

[1]School of Exercise and Nutritional Sciences, San Diego State University, San Diego, CA; [2]Exercise and Physical Activity Resource Center, University of California, San Diego, San Diego, CA; and [3]Center for Wireless and Population Health Systems, University of California, San Diego, San Diego, CA

ABSTRACT

KLEPIN, K., D. WING, M. HIGGINS, J. NICHOLS, and J. G. GODINO. Validity of Cardiorespiratory Fitness Measured with Fitbit Compared to $\dot{V}O_{2max}$. *Med. Sci. Sports Exerc.*, Vol. 51, No. 11, pp. 2251–2256, 2019. **Purpose:** Cardiorespiratory fitness (CRF), broadly defined as the body's ability to utilize oxygen, is a well-established prognostic marker of health, but it is not routinely measured. This may be due to the difficulty of acquiring high-quality CRF measures. The purpose of this study was to independently determine the validity of the Fitbit Charge 2's measure of CRF (Fitbit CRF). **Methods:** Sixty-five healthy adults between the ages of 18 and 45 yr (55% female, 45% male) were recruited to undergo gold standard $\dot{V}O_{2max}$ testing and wear a Fitbit Charge 2 continuously for 1 wk during which they were instructed to complete a qualifying outdoor run to derive the Fitbit CRF (units: mL·kg^{-1}·min^{-1}). This measure was compared with $\dot{V}O_{2max}$ measures (units: mL·kg^{-1}·min^{-1}) epoched at 15 and 60 s. **Results:** Bland–Altman analyses revealed that Fitbit CRF had a positive bias of 1.59 mL·kg^{-1}·min^{-1} compared with laboratory data epoched at 15 s and 0.30 mL·kg^{-1}·min^{-1} compared with data epoched at 60 s ($n = 60$). F statistics (2.09; 0.08) and P values (0.133; 0.926) from Bradley–Blackwood tests for the concordance of Fitbit CRF with 15- and 60-s laboratory data, respectively, support the null hypothesis of equal means and variances, indicating there is concordance between the two measures. Mean absolute percentage error was less than 10% for each comparison. **Conclusions:** The Fitbit Charge 2 provides an acceptable level of validity when measuring CRF in young, healthy, and fit adults who are able to run. Further research is required to determine if it is a potentially useful tool in clinical practice and epidemiological research to quantify, categorize, and longitudinally track risk for adverse outcomes. **Key Words:** CARDIORESPIRATORY FITNESS, $\dot{V}O_{2max}$, GRADED EXERCISE TEST, ACTIVITY TRACKER

The epidemics of obesity, diabetes, and cardiovascular disease are now global in scale (1), and both their incidence and prevalence are expected to increase as a result of the aging of the population and an exacerbation of health disparities (2). The risk for these common, complex chronic diseases and their associated comorbidities can be substantially reduced through improvements in cardiorespiratory fitness (CRF) (3,4). Cardiorespiratory fitness, broadly defined as the body's ability to transport, absorb, and utilize oxygen is a well-established prognostic marker of health (3–6). In fact,

there is increasing epidemiological and clinical evidence that suggests that CRF may be a stronger predictor of all-cause mortality than other chronic disease risk factors, such as smoking, hypertension, high cholesterol, and type 2 diabetes (4,7,8). Although CRF has been shown to significantly improve the reclassification of risk for adverse outcomes (9–12), it is not routinely measured (3).

This may be due, at least in part, to the difficulty of acquiring high-quality CRF measures. The "gold standard" measure of CRF is maximal oxygen uptake, or $\dot{V}O_{2max}$, which is assessed during a graded exercise test typically conducted on a treadmill or cycling ergometer (3,13,14). This requires individuals to wear a face-mask that enables the measurement of breath-by-breath volume and fractional composition of inspired and expired gases. This type of CRF test not only requires substantial engagement by the individual being tested but also significant expertise, time, and cost to implement, making it impractical in most clinical and epidemiological contexts. A somewhat less burdensome measure of CRF can be derived from a 12-min run test (also known as a "Cooper Test"), which requires individuals to run as far as possible for up to 12 min on a flat course (15,16). $\dot{V}O_{2max}$ is then estimated from the total distance traveled according to well-established age- and sex-based population norms (15,16). Although this test requires less expertise, time, and cost to

APPLIED SCIENCES

conduct than a $\dot{V}O_{2max}$ test, it too may not be feasible to widely undertake.

Recent advances in microtechnology, data processing, wireless communication, and battery capacity have resulted in the proliferation of low-cost, noninvasive wearable devices that seamlessly integrate with the wearer's smartphone and can be used to measure multiple health-related signals in a free-living environment (17). One such device is the Fitbit Charge 2, a low-cost wrist-worn activity tracker (Fitbit Inc., San Francisco, CA, https://www.fitbit.com/charge2). Among other things, it contains a triaxial accelerometer, an optical heart rate monitor, and an altimeter. When linked with the GPS sensor on a wearer's smartphone during an outdoor run on flat terrain at a comfortable pace that lasts at least 10 min, Fitbit will utilize the wearer's heart rate and pace during the run, along with their resting heart rate, age, sex, and weight to calculate an estimate of CRF (the exact algorithm used is proprietary and currently unknown). Like the aforementioned 12-min run test, this method relies heavily on a structured run of a known duration, suggesting a great deal of face validity. However, the test validity of the Fitbit Charge 2's measure of CRF has not been investigated to date.

In the present study, we assessed the test validity of the Fitbit Charge 2's measure of CRF by comparing it with $\dot{V}O_{2max}$ measured during a graded exercise test conducted on a treadmill using state-of-the-science equipment. This study represents a logical step toward being able to make an informed decision about whether or not the Fitbit Charge 2's measure of CRF could be used within clinical practice and epidemiological research. Given that CRF is a very informative marker of overall health, the potential to accurately and cheaply measure it via a consumer-level wearable in a free-living environment has important implications for its widespread adoption.

METHODS

Participants. Potential participants were recruited via a combination of print (e.g., flyers) and digital (e.g., email) advertisements. Eligible participants were adults age 18 to 45 yr, free from chronic diseases or injuries that would impede the completion of a graded-exercise test to volitional fatigue and at least three outdoor runs of 15 min or more, owned a smartphone capable of running the Fitbit application and pairing to the Fitbit Charge 2 with GPS enabled, and spoke English. Potential participants were excluded if they answered affirmatively to one or more questions in the American College of Sports Medicine's Physical Activity Readiness Questionnaire (18), indicated that they could not run continuously for at least 15 min without stopping, or indicated they were pregnant.

Procedures and measures. All study procedures were approved by the University of California, San Diego Institutional Review Board (approval number 161732). All participants provided written informed consent and attended two in-person study visits at the Exercise and Physical Activity Resource Center (EPARC).

During the first visit, participants self-reported sex and age, and EPARC staff measured participants' weight (to the nearest 0.1 kg) and height (to the nearest 0.1 cm) using a calibrated digital scale and stadiometer (Seca, Chino, CA). Both weight and height were measured with participants wearing lightweight clothes but without shoes, and two separate measurements were averaged (if weight or height measurements differed by more than 1%, then a third measure was taken and the average of the two measures that differed by less than 0.02 kg or 0.05 cm, respectively, was taken). Body mass index was calculated as weight in kilograms divided by height in square meters.

Participants then completed a maximal graded exercise test on a Quinton Q-Stress treadmill (Mortara, Milwaukee, WI) that was calibrated monthly for accuracy of speed and grade. The maximal graded exercise test protocol began with a warm-up at a self-selected pace on the treadmill for 5 to 10 min. During the warm-up, EPARC staff explained how to use the Borg RPE and reminded participants that they were expected to achieve their maximal level of exertion. Participants were then equipped with a breath mask that covers the nose and mouth (KORR Medical Technologies, Salt Lake City, UT), and a Bluetooth enabled heart rate monitor worn on the chest (Garmin, Olathe, KS). The preprogrammed treadmill protocol began with participants running at 5 mph (5.0 mph) with 0% incline for 3 min (13,19–21). The workload was then increased approximately 0.75 METs every minute (13,19–21). This was achieved via an increase in speed (0.5 mph·min^{-1}) for the first 2 min, and an increase in incline by 1.5% every minute thereafter (13,19–21). RPE was assessed during the final 10 s of each minute, and the protocol continued until the participant signaled to stop (i.e., indication of volitional fatigue) (13,19–21). Upon indication of volitional fatigue, the treadmill was immediately slowed to 2.0 mph, and participants were encouraged to walk until completely recovered. Breath by breath oxygen uptake ($\dot{V}O_2$) was continuously measured using an indirect calorimeter (COSMED, Trentino, Italy) that was calibrated for gas volume and fractional composition immediately (i.e., less than 30 min) before the start of the maximal graded exercise test protocol. At present, there is no consensus on the length of the epoch to use when averaging breath-by-breath level $\dot{V}O_2$ data, but there is evidence that void of steady state $\dot{V}O_2$ consumption, shorter epochs are more likely to elicit higher values (15,20,22). The extent to which higher values are more accurate remains unclear. Therefore, to present a range of epochs likely to be used, $\dot{V}O_2$ data were averaged into 15- and 60-s epochs, and the largest value recorded during these epochs was identified as $\dot{V}O_{2max}$ in analyses (i.e., 15-s CRF and 60-s CRF) (15,20,22). Use of indirect calorimetry is the gold standard method for assessing CRF (3,13,19–21).

EPARC staff also downloaded the Fitbit application onto participants' smartphone and logged into a study-specific Fitbit account that was created using a unique username and password (i.e., the participant was not identified), and paired each participant's phone to a study provided Fitbit Charge 2. The study-specific account was then populated with each

participant's age, sex, handedness, and measured height and weight. EPARC staff explained how to properly wear the Fitbit Charge 2 and use it for GPS tracked outdoor runs. Participants were instructed to complete at least three GPS tracked outdoor runs on flat terrain at a comfortable pace lasting at least 15 min over the following week. They were also instructed to wear the device continuously except while swimming or bathing. A pamphlet detailing this information was provided to each participant. After the establishment of a resting heart rate and a qualifying run, Fitbit utilized a participant's heart rate and pace during the run, along with their resting heart rate, age, sex, and weight to calculate an estimate of CRF. The exact algorithm used is proprietary and currently unknown.

During the second visit, which occurred approximately 1 wk after the first, EPARC staff manually recorded participants' CRF as calculated by Fitbit (i.e., Fitbit CRF). The Fitbit Charge 2 was then unpaired from the participant's phone, and the Fitbit account was closed. Participants were also asked to complete a widely utilized system usability scale questionnaire asking about the intuitiveness of the Fitbit Charge 2 and corresponding smart phone application (23), and whether they believed that the device and application would be helpful in improving physical fitness. Questions were rated on a five-point Likert scale ranging from strongly disagree (1) to strongly agree (5). Scores were recalculated on a 0- to 4-scale, summed, and multiplied by 2.5 to create a 100-point scale with higher scores indicating higher usability (23,24). As compensation for completion of the study, participants were given a feedback report about their $\dot{V}O_{2max}$, lactate threshold, and potential training zones.

Statistical analysis. Demographic and anthropometric characteristics of the study sample were described using univariate descriptive statistics (i.e., proportions and means and standard deviations). Test validity was described using Bland–Altman procedures to analyze the agreement of 15-s CRF and 60-s CRF with Fitbit CRF. (25). Bradley–Blackwood tests were used for a simultaneous analysis of the concordance between means and variances of the respective measures (26). Mean absolute percentage error was calculated as the average of absolute differences between the measures, divided by the relevant $\dot{V}O_{2max}$, multiplied by 100. CRF measures were categorized according to age- and sex-based population norms defined as superior, excellent, good, fair, and poor (27). Categories were also collapsed into groups defined as superior or excellent, good, and fair or poor, because these categories are aligned with those used in risk stratification for all-cause mortality (6). The binary agreement between the aforementioned categories was analyzed using χ^2 tests of independence. All statistical analyses were conducted using STATA 13.0 (StataCorp, College Station, TX).

RESULTS

From June 4, 2017, to December 4, 2017, 65 participants enrolled in the study. One participant experienced an equipment malfunction during the maximal graded exercise test and did not continue in the study. Another voluntarily dropped out before completing all measures. Three participants did not complete a GPS tracked outdoor run that allowed for the calculation of Fitbit CRF. A total of 60 participants (27 male and 33 females) completed all study protocols and were included in data analyses. The mean (SD) age was 31.0 yr (7.3 yr), mean (SD) height was 169.5 cm (10.5 cm), mean (SD) weight was 70.2 kg (14.1 kg), and mean (SD) body mass index was 24.3 kg·m^{-2} (3.3 kg·m^{-2}) (Table 1).

Figure 1 shows that when compared to 15-s CRF, Fitbit CRF had a positive mean bias of 1.59 mL·kg^{-1}·min^{-1} with upper and lower limits of 13.28 and −10.10, respectively. Compared with 60-s CRF, Fitbit CRF had a positive mean bias of 0.30 mL·kg^{-1}·min^{-1} with upper and lower limits of 11.96 and −11.36, respectively. For each comparison, the F statistic (15-s CRF vs Fitbit CRF = 2.09; 60-s CRF vs Fitbit CRF = 0.08) and corresponding P value (15-s CRF vs Fitbit CRF = 0.133; 60-s CRF vs Fitbit CRF = 0.926) of the Bradley–Blackwood test supports the null hypothesis of equal means and variances indicating that there is concordance between measures regardless of the epoch used in the gold standard. The Bland–Altman plots also revealed two observations that fell outside the limits of agreement (3.3%) within each comparison. The mean absolute percentage error was nearly equal when Fitbit CRF was compared with 15-s CRF and 60-s CRF, with values of 9.41% and 9.14%, respectively.

Figure 2 shows that Fitbit CRF correctly classified category of fitness 70.00% (42/60) of the time when compared with both 15-s CRF and 60-s CRF. These estimates improved when categories are binned as superior or excellent, good, fair or poor, with 91.70% (55/60) with both 15-sCRF and 60-s CRF. For each comparison, the χ^2 statistic (15 c CRF vs Fitbit CRF = 66.93; 60-s CRF vs Fitbit CRF = 64.33) and corresponding P value (both <0.001) reject the null hypothesis of independence, indicating an association between the measures.

Three participants who completed all of the physical assessments did not complete the system usability scale, thus data from 57 participants were analyzed. The mean (SD) score in reference to the Fitbit Charge 2 was 79.8 (15.1),

TABLE 1. Participant characteristics.

N	60 (100)
Age (yr)	
18 to 25	13 (21.7)
26 to 30	20 (33.3)
31 to 35	9 (15.0)
36 to 40	9 (15.0)
41 to 45	9 (15.0)
Sex	
Female	33 (55)
Male	27 (45)
Height (cm), mean (SD)	169.51 (11.03)
Weight (kg), mean (SD)	70.29 (15.28)
Body mass index (kg·m^{-2}), mean (SD)	24
CRF	
15-s CRF (mL·kg^{-1}·min^{-1}), mean (SD)	48.9 (8.2)
60-s CRF (mL·kg^{-1}·min^{-1}), mean (SD)	47.6 (8.1)
Fitbit CRF (mL·kg^{-1}·min^{-1}), mean (SD)	47.3 (8.1)

*Values are numbers (percentages) unless otherwise specified.

APPLIED SCIENCES

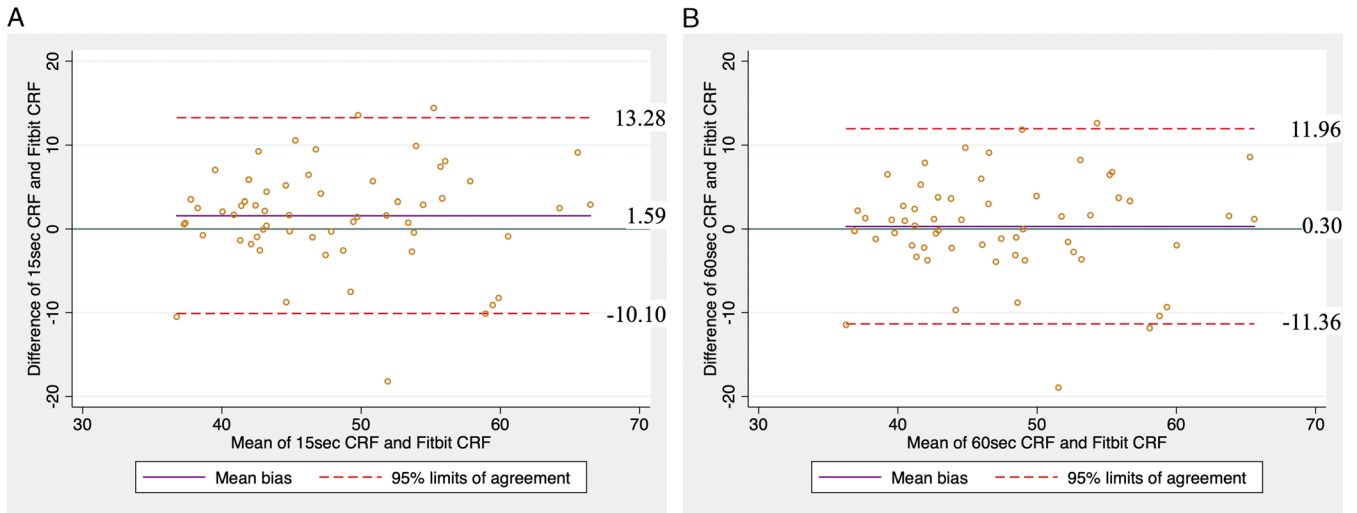**FIGURE 1—Bland–Altman plots showing the agreement Fitbit CRF with (A) 15-s CRF and (B) 60-s CRF. Unit of measure, mL·kg^{-1}·min^{-1}.**

and the mean (SD) score in reference to the corresponding smartphone application was 80.9 (12.5). These scores correspond to an adjective rating of "excellent" acceptability (24). Additionally, when asked if information from the Fitbit Charge 2 and corresponding application would motivate them to be more active over the long-term, most participants agreed (mean [SD], 4.2 [0.87]).

DISCUSSION

The aim of this study was to assess the test validity of the Fitbit Charge 2's measure of CRF when compared with the current gold standard measures of $\dot{V}O_{2max}$ assessed using indirect calorimetry in a healthy population. By collecting breath-by-breath data and averaging across multiple possible epochs, we were able to examine this agreement at several potentially meaningful levels. Specifically, we analyzed the validity of Fitbit CRF against "true" maximal capacity which is likely observed when small changes in oxygen uptake are averaged over short epochs (e.g., 15 s), and also against longer epochs (e.g., 60 s) like those utilized for generating predictive algorithms in commonly utilized field assessments of CRF.

Regardless of the epoch used, there was a significant association between Fitbit CRF and $\dot{V}O_{2max}$, although agreement improved when 60-s epochs were used in the laboratory-based measure. With an average bias of only 0.3 mL·kg^{-1}·min^{-1} over minute level epochs and a mean absolute error less than 10%, the Fitbit Charge 2 provides an acceptable level of validity when measuring CRF. As such, it appears that the Fitbit Charge 2 offers many of the benefits implicit in submaximal field testing (i.e., lower cost, less risk of injury, etc.). Additionally, because the device can be worn over long periods, there is an added opportunity for free-living, longitudinal tracking of CRF.

Although specific $\dot{V}O_{2max}$ values can be useful for targeted physical training, their clinical and epidemiological use is magnified when used for risk stratification. It is here that Fitbit CRF may have an important impact. The χ^2 analysis indicated statistically significant high categorical agreement (70.0%) when five levels of fitness were utilized. When further collapsed to three categories, more in-line with the risk stratification proposed by Blair et al. (6), agreement was high (91.7%). Importantly, these findings are perhaps unsurprising given that such a large proportion of the sample in the present study was classified as having a fitness level that was either superior or
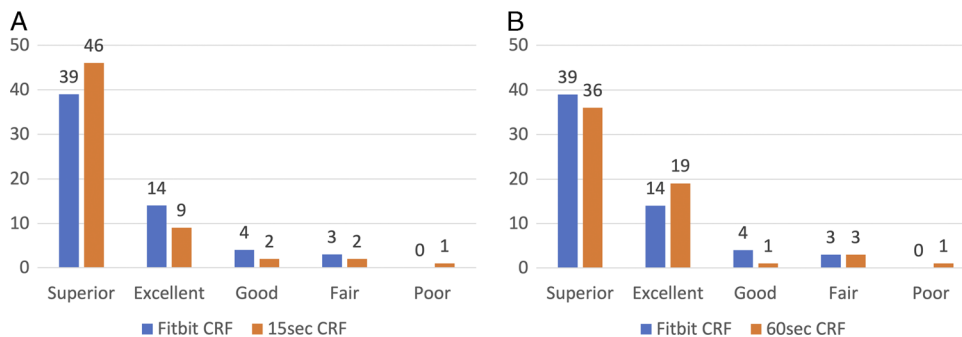

FIGURE 2—Categorical comparison of Fitbit CRF with (A) 15-s CRF and (B) 60-s CRF. Unit of measure, mL·kg^{-1}·min^{-1} placed into age- and sex-based population norms.

excellent. Before strong conclusions about these findings can be made, additional research including populations with low levels of fitness and chronic diseases are necessary to more robustly determine if Fitbit CRF can be used to quantify, categorize, and longitudinally track risk for adverse outcomes.

Results from participants' responses on the usability and acceptability of the Fitbit Charge 2 and corresponding smartphone application are promising for the prospect of widespread adoption in free-living populations. Specifically, participants found both the device and smartphone application easy to use and potentially helpful in regard to motivating healthy levels of physical activity. If the results of this study are replicated in more clinically relevant populations (i.e., those with low fitness levels and chronic disease), then Fitbit may provide a platform for relatively inexpensive collection of large-scale, longitudinal data regarding CRF.

Although the data gathered in this study are promising, the findings should be considered within its limitations. First, the majority of participants had a high fitness level and were able to run. Further research is needed to determine if Fitbit CRF can be accurately derived when individuals transition from running to walking, or while walking throughout the entirety of an assessment. Additionally, we recruited a relatively young sample that likely had a high level of familiarity and comfort with mobile technology in general, and smartphone-based applications in particular. Additional research is necessary to determine if the Fitbit Charge 2 provides valid measures of CRF in a heterogeneous sample with lower overall fitness, greater age, existing disease, and less confidence in the use of mobile technology. An additional limitation is that participants may have arrived at volitional fatigue before achieving their "true" maximal capacity during the laboratory-based measurement. Lastly, all research in which a single measurement is used as a "gold standard" is susceptible to random error, and in this case, it is impossible to know how that error influenced the estimates of bias.

CONCLUSIONS

The Fitbit Charge 2's measure of CRF offers an acceptably valid estimate of $\dot{V}O_{2max}$ in a young, healthy, and fit population of adults who were able to run. This free-living measure of CRF can be assessed at relatively low cost and with a relatively high level of acceptability. As such, it appears that the Fitbit Charge 2 offers many of the benefits implicit in submaximal field testing, and because the device can be worn over long periods, it presents an added opportunity for free-living, longitudinal tracking of CRF. Additional research is needed to determine if these results can be replicated in more clinically relevant populations.

REFERENCES

1. Murray CJ, Vos T, Lozano R, et al. Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet*. 2012;380(9859):2197–223.
2. Bauer UE, Briss PA, Goodman RA, Bowman BA. Prevention of chronic disease in the 21st century: elimination of the leading preventable causes of premature death and disability in the USA. *Lancet*. 2014;384(9937):45–52.
3. Ross R, Blair SN, Arena R, 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):e653–99.
4. 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 Netw Open*. 2018;1(6):e183605.
5. Myers J, McAuley P, Lavie CJ, Despres JP, Arena R, Kokkinos P. Physical activity and cardiorespiratory fitness as major markers of cardiovascular risk: their independent and interwoven importance to health status. *Prog Cardiovasc Dis*. 2015;57(4):306–14.
6. Blair SN, Kohl HW 3rd, Paffenbarger RS Jr, Clark DG, Cooper KH, Gibbons LW. Physical fitness and all-cause mortality. A prospective study of healthy men and women. *JAMA*. 1989;262:2395–401.
7. Laukkanen JA, Rauramaa R, Salonen JT, Kurl S. The predictive value of cardiorespiratory fitness combined with coronary risk evaluation and the risk of cardiovascular and all-cause death. *J Intern Med*. 2007;262(2):263–72.
8. Kokkinos P. History of physical activity and health. In: *Physical Activity and Cardiovascular Disease Preventions*. London: Jones and Barlett Publishers; 2010. pp. 3–18.
9. Stamatakis E, Hamer M, O'Donovan G, Batty GD, Kivimaki M. A non-exercise testing method for estimating cardiorespiratory fitness: associations with all-cause and cardiovascular mortality in a pooled analysis of eight population-based cohorts. *Eur Heart J*. 2013;34(10):750–8.
10. Gupta S, Rohatgi A, Ayers CR, et al. Cardiorespiratory fitness and classification of risk of cardiovascular disease mortality. *Circulation*. 2011;123(13):1377–83.
11. 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.
12. Chang P, Nead KT, Olin JW, Myers J, Cooke JP, Leeper NJ. Effect of physical activity assessment on prognostication for peripheral artery disease and mortality. *Mayo Clin Proc*. 2015;90(3):339–45.
13. American College of Sports Medicine. *ACSM's Guidelines for Exercise Testing and Prescription*. 9th ed. Philadelphia, PA: Wolters Kluwer/Lippincott Williams & Wilkins Health; 2014. pp. 1–480.
14. Balady GJ, Arena R, Sietsema K, et al. Clinician's guide to cardiopulmonary exercise testing in adults: a scientific statement from the American Heart Association. *Circulation*. 2010;122(2):191–225.
15. Cooper KH. A means of assessing maximal oxygen intake. Correlation between field and treadmill testing. *JAMA*. 1968;203(3): 201–4.

APPLIED SCIENCES

16. Noonan V, Dean E. Submaximal exercise testing: clinical application and interpretation. *Phys Ther*. 2000;80(8):782–807.

17. Chan M, Estève D, Fourniols JY, Escriba C, Campo E. Smart wearable systems: current status and future challenges. *Artif Intell Med*. 2012;56(3):137–56.

18. Adams R. Revised physical activity readiness questionnaire. *Can Fam Physician*. 1999;45:992, 995, 1004–5.

19. Shephard RJ, Allen C, Benade AJ, et al. The maximum oxygen intake. An international reference standard of cardiorespiratory fitness. *Bull World Health Organ*. 1968;38(5):757–64.

20. Fletcher GF, Ades PA, Kligfield P, et al. Exercise standards for testing and training: a scientific statement from the American Heart Association. *Circulation*. 2013;128(8):873–934.

21. 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;2016:1–12.

22. Astorino TA. Alterations in VO2 max and the VO2 plateau with manipulation of sampling interval. *Clin Physiol Funct Imaging*. 2009;29(1):60–7.

23. Brooke J. SUS—a quick and dirty usability scale. *Usability Eval Ind*. 1996;189(194):4–7.

24. Bangor A, Staff T, Kortum P, Miller J. Determining what individual SUS scores mean. *J usability Stud*. 2009;4(3):114–23.

25. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*. 1986;1(8476):307–10.

26. Bradley EL, Blackwood LG. Comparing paired data: a simultaneous test for means and variances. *Am Stat*. 1989;43(4):234–5.

27. Heyward VH. The physical fitness specialist certification manual. In: *Advance Fitness Assessment & Exercise Prescription*. 3rd ed. Dallas, TX: The Cooper Institute for Aerobics Research; 1997. p. 48.

28. Braddock CH 3rd, Snyder L. The doctor will see you shortly. *J Gen Intern Med*. 2005;20(11):1057–62.