Test–retest repeatability of cardiopulmonary exercise test variables in patients with cardiac or respiratory disease
Anthony Barron, Niti Dhutia, Jamil Mayet, Alun D Hughes, Darrel P Francis and Roland Wensel
European Journal of Preventive Cardiology published online 7 January 2014
DOI: 10.1177/2047487313518474

The online version of this article can be found at:
http://cpr.sagepub.com/content/early/2014/01/07/2047487313518474
A more recent version of this article was published on - Mar 18, 2014
Test–retest repeatability of cardiopulmonary exercise test variables in patients with cardiac or respiratory disease

Anthony Barron¹, Niti Dhutia¹, Jamil Mayet¹, Alun D Hughes¹, Darrel P Francis¹ and Roland Wensel¹,²

Abstract

Background: The test–retest reliability for multiple cardiopulmonary exercise test (CPX) variables has not been compared in a single study and the influence of different diseases on test–retest reliability has not been examined. We investigated different measures of test–retest reliability for multiple variables and compared them by category of cardiac or respiratory disease.

Methods: Patients with chronic obstructive airways disease (n = 24), heart failure (n = 43), or severe mitral valve disease (n = 26) were recruited into a prospective study. Each patient underwent two bicycle ergometer tests; the first, a familiarization test, with a 10 W/min ramp, and the second a personalized ramp based on the results of the familiarization test to elicit maximal effort within 8–10 min. Intraclass correlation coefficients (ICC) and coefficients of variation between the two tests were calculated. Influence of potential modifiers was assessed using repeated measures analysis of variance.

Results: Peak VO₂ (ICC 0.95, 95% CI 0.94–0.97), oxygen uptake efficiency slope (ICC 0.93, 95% CI 0.90–0.95), O₂ pulse (ICC 0.96, 95% CI 0.94–0.97), and the VE/VCO₂ ratio at the nadir (ICC 0.92, 95% CI 0.89–0.95) all showed excellent test–retest reliability, with within-subject coefficients of variation <0.12. VO₂ at the anaerobic threshold (ICC 0.84, 95% CI 0.78–0.89) and the VE/VCO₂ slope (ICC 0.88, 95% CI 0.79–0.93) showed good test–retest reliability, although inferior to peak VO₂. Age, gender, body mass index, disease aetiology, protocol change, and intertest interval did not affect the reliability of most variables.

Conclusions: CPX showed high test–retest reliability; certain variables such as peak VO₂ and oxygen uptake efficiency slope outperform others. These results identify which variables are most suitable for serial testing of patients with three common disease aetiologies owing to their superior reproducibility.

Keywords
Cardiopulmonary exercise testing, reproducibility, test–retest reliability

Received 26 July 2013; accepted 8 December 2013

Introduction

A test of a chronic feature of the patient should, ideally, give the same value every time it is conducted. Factors affecting a test’s reproducibility can be broken down into two groups: random error, which includes within-patient biological and temporal variations and device/operator/reporter variability; and systematic error, which includes phenomena such as familiarization or fatigue which cause results to be consistently under or overestimated within a particular individual.¹

¹Imperial College, London, UK
²West Hertfordshire Hospitals NHS Trust, UK

Corresponding author:
Anthony Barron, International Centre for Circulatory Health, Imperial College, 59 North Wharf Road, London W2 1LA, UK.
Email: anthonybarron@hotmail.co.uk
gas-exchange variables such as oxygen consumption (VO₂) and carbon dioxide production (VCO₂) is around 0.01–0.03%, and for flow variables such as tidal volume and respiratory frequency around 1–2%. The majority of derived variables are calculated using computer software, minimizing interpretation error. The majority of variation between two tests in the same patient will be due to temporal and biological variation and technical imperfections such as mask leak.

In research, participants are normally tested on standard protocols, yet in clinical practice, protocols are often not standardized. Research studies often adhere to strict protocols rarely enforced in routine clinical practice, for example, the fasting state. Therefore a number of questions remain regarding the reproducibility of CPX variables.

How do the test–retest reliabilities of each variable compare with one another in the same study under typical clinical rather than ideal research conditions? Does disease aetiology affect reproducibility? How does protocol affect reproducibility of the test? How does intertest interval affect reproducibility?

Measurements of slopes (VE/VCO₂ slope and oxygen uptake efficiency slope) are considered to be more reproducible than instantaneous or averaged data variables such as peak VO₂ or VE/VCO₂ ratio. However, this issue has not been examined formally.

We assessed, for typical patients with chronic obstructive pulmonary disease (COPD), chronic heart failure (CHF), and mitral valve disease, the test–retest reliability of multiple CPX variables, and the influence of disease aetiology, protocol changes, and intertest time interval on the test–retest reliability.

**Methods**

**Subjects**

Participants were recruited from Imperial College Healthcare NHS Trust or the Royal Brompton and Harefield Hospital NHS Trust, in London, UK. Patients with CHF, mitral valve disease requiring surgical intervention, or COPD underwent CPX. B-type natriuretic peptide measurement and full lung function testing. CHF patients must have been previously symptomatic but currently stable and the majority either the prior recipients of or awaiting cardiac resynchronization therapy. Appropriate consecutive patients awaiting surgical repair/replacement of their mitral valve were approached but were not necessarily symptomatic; surgery was indicated according to current guidelines.

Principal exclusion criteria included the inability to perform a symptom-limited exercise test, significant renal impairment (estimated glomerular filtration rate <30 ml/min/1.73 m²), symptomatic coronary artery disease, recent acute cardiorespiratory illness/decompensation, anaemia, morbid obesity, and all standard contraindications to exercise testing. The study conformed to the principles of the Declaration of Helsinki and was approved by the Research Ethics Committee of Imperial College Healthcare NHS Trust. All participants provided written informed consent.

**Lung function and cardiopulmonary exercise testing**

Full lung function testing was performed on a SpiroAir (Medisoft, Sorinnes, Belgium), including spirometry, diffusion, and lung volume subdivisions.

Patients underwent exercise testing on an ergoselect 100 ergometer (Ergoline, Baden-Württemberg, Germany) in an air-conditioned room after equipment familiarization. Patients were encouraged to abstain from caffeine for 12 h prior. Only small meals were encouraged, with at least a 2-h interval before testing. Spirometry and exercise testing were performed using a Quark CPX System (COSMED, Rome, Italy), calibrated before each test. The test comprised 3 min of rest followed by 3 min of unloaded cycling before a 10 W/min ramp, exercising to symptom-limited exhaustion with standard encouragement, followed by active recovery. Talking was discouraged, unless clinically necessary. Gas exchange was monitored breath-by-breath. Blood pressure was recorded every 3 min by manual sphygmomanometry.

At least 2 h later, and sometimes on a separate day, patients underwent a second CPX. Similarly to the first test, there were 3-min rest, 3-min unloaded and active recovery periods, but with a ramp protocol based on the results of the initial test, aiming to elicit exhaustion at 8–10 min of incremental exercise. The breakdown of frequencies of each ramp protocol is shown in Supplementary Figure S1 (available online). Data were anonymized and analysis performed blind to identity and test sequence, although most of the variables (with the exception of the identification of ventilatory thresholds) were calculated by a formal prespecified trial analysis protocol and therefore should not be susceptible to human preference.

**Variables**

For full information on CPX variable calculations, see the online supplement. Briefly, all peak variables were measured using the highest 20-s average within the final minute of exercise and the first 10 s of recovery. All other variables were measured using standard methods. The VE/VCO₂ slope was calculated in two ways: slope 1 used all data until the ventilatory compensation point (VCP); slope 2 used all exercise data (i.e. including post-VCP data). Maximal voluntary ventilation was
calculated as forced expiratory volume 1 (FEV1) × 40. Exercise-induced shunting and exercise oscillatory ventilation were determined using previously described methodology.10,11

Statistical analysis

Statistical analysis was performed using Stata version 11.1 for Windows (StataCorp, College Station, TX, USA). All patient characteristics and CPX variables were assessed for normality using Q-Q plots and Shapiro–Wilk tests, and logarithmic transformation if appropriate. One-way analysis of variance (with age, gender, and body mass index (BMI) as covariates) was used to compare baseline characteristics between disease categories. Age, gender, and BMI were tested without the addition of covariates. Test–retest data were examined using Bland–Altman plots and the mean difference and standard deviation of the difference (SDD) calculated. The within-subject coefficient of variation (CoV) was calculated as the ratio of the SDD and the mean of each variable. Intraclass correlation coefficients (ICC) were calculated for the pooled group and each disease category, with comparison performed by bootstrapping. We analysed six potential modifiers of reproducibility using mixed linear models: age (separated above and below median), gender, weight status (BMI ≤ 25 kg/m²), aetiology of disease, whether tests were performed on the same day (intertest interval) and whether the ramp protocol was steeper, shallower, or the same as the original. These final two modifiers were not randomized, with intertest interval determined by the subject, and second ramp protocol determined by achieved work rate on test 1. A significant interaction term (p < 0.05) identified that the difference in a variable between tests 1 and 2 was influenced by that factor. p < 0.05 was considered statistically significant.

Results

A total of 100 patients were recruited into the study, of which seven were excluded. Of the final 93 (70 male), 43 were patients with CHF, 26 with mitral valve disease, and 24 with COPD. The patient characteristics of the total group and by disease category are shown in Table 1. Of the 26 patients with mitral valve disease, 18 were symptomatic. The protocols used for test 2 included 6 W/min (n = 7), 8 W/min (n = 12), 10 W/min (n = 31), 12 W/min (n = 18), 15 W/min (n = 21), and 20 W/min (n = 4).

There were no significant differences in age, height, prevalence of hypertension, diabetes, or ethnicity (p > 0.05) between disease groups. There were significantly less women in the CHF group (p = 0.02).

Weight (p = 0.01) was significantly higher in the CHF group but not after adjustment for age and gender (p = 0.06). Table 1 shows other between-group differences.

There was a nonsignificant trend to shorter exercise times within COPD patients and longer times within the mitral group. The protocol chosen for test 2 reflected this difference in test 1 time, with a nonsignificant trend to the highest average protocol chosen in the mitral valve disease group and lowest in the COPD group.

Despite 15 patients desaturating during exercise (fall in arterial oxygen saturations ≥4%), none displayed criteria diagnostic for exercise-induced shunting. There were 49 patients who had a noticeable inflection point of the VE/VCO₂ relationship at the VCP on at least one test.

Difference between test 1 and test 2

Table 2 shows the overall means, mean differences and 95% limits of agreement for test 1 and test 2 for 35 CPX variables. Peak VO₂, anaerobic threshold (AT), peak heart rate (HR), double product (DP), peak minute ventilation and respiratory frequency, and peak work rate showed significant positive differences between tests 1 and 2. End-tidal CO₂ (pETCO₂) at the AT and breathing reserve at peak showed a significant reduction between tests 1 and 2. Peak respiratory exchange ratio (RER) did not differ between tests suggesting that effort was similar. Oxygen uptake efficiency slope (OUES), oxygen uptake efficiency plateau (OUEP), O₂ pulse, the VO₂/WR slope, and all measures of the VE/VCO₂ relationship were not significantly different between tests.

Bland–Altman plots for some of the principal variables show the differences between tests 1 and 2 (Figure 1).

Test–retest reliability of various CPX variables

Table 2 shows the ICC, SDD, and the CoV for each variable. Peak VO₂, OUES measured using full data, OUEP, O₂ pulse, peak circulatory power, the VE/VCO₂ ratio at the VCP and nadir, and breathing reserve at the AT and peak all showed excellent test–retest reliability, with an ICC ≥ 0.90. The RER at rest and AT and resting arterial O₂ saturations performed less well (ICC ≤ 0.69). All other measures showed more intermediate values of test–retest reliability.

The VE/VCO₂ ratio measured at nadir was significantly more reproducible than the ratio measured at the AT. The slopes of the VE/VCO₂ relationship using full data or data up to the VCP showed similar test–retest reliability.
Table 1. Patient characteristics

|                     | All                  | COPD  
| (n = 24)           | CHF  
| (n = 43)           | MVD  
| (n = 26)  | p between  
| groups       |
| Age (years)        | 65.2 ± 10.9          | 66.6 ± 9.5 |
|                   | 66.6 ± 11.1          | 61.7 ± 11.3 |
|                   | 1.06                 |           |
| Male*              | 70 (75.3)            | 16 (66.7) |
|                   | 38 (88.4)            | 16 (61.5) |
|                   | 0.02                 |           |
| BMI (kg/m²)*       | 26.7 ± 4.4           | 25.8 ± 4.6 |
|                   | 28.1 ± 4.1           | 25.3 ± 4.0 |
|                   | 0.02                 |           |
| Height (cm)        | 170.1 ± 10.0         | 166.8 ± 7.8 |
|                   | 171.3 ± 9.2          | 171.2 ± 12.3 |
|                   | 0.18                 |           |
| Weight (kg)        | 77.6 ± 15.8          | 72.0 ± 16.1 |
|                   | 82.7 ± 15.5          | 74.4 ± 13.8 |
|                   | 0.24                 |           |
| Diabetics          | 10 (10.7)            | 1 (4.2)   |
|                   | 8 (18.6)             | 1 (3.8)   |
|                   | 0.08                 |           |
| Current smokers    | 12 (12.9)            | 5 (20.8)  |
|                   | 7 (16.3)             | 0 (0)     |
|                   | 0.06                 |           |
| Hypertensive       | 40 (43.0)            | 10 (41.7) |
|                   | 19 (44.2)            | 11 (42.3) |
|                   | 0.42                 |           |
| B-type natriuretic peptide (pg/ml)* | 86 (37–203) | 36 (14–52) |
|                   | 118 (85–286)         | 85 (43–222) |

Values are mean ± SD, n (%), or median (interquartile range). Differences between groups by ANOVA (age, gender, and BMI as covariates) for continuous variables and chi-squared analysis for categorical variables; age, gender, and BMI are performed on a model without covariates. ANOVA on FVC, FEV1/FVC ratio, and resting heart rate were calculated following logarithmic transformation but raw data are presented as median (interquartile range); *p > 0.05 between CHF and mitral valve disease; **p > 0.05 between COPD and CHF; ***p > 0.05 between COPD and mitral valve disease; AT, anaerobic threshold; BMI, body mass index; CHF, chronic heart failure; COPD, chronic obstructive pulmonary disease; DLCO, diffusing capacity of the lung for carbon monoxide; FEV, forced expiratory volume; FVC, forced vital capacity; MVD, mitral valve disease.

OUES had significantly better test–retest reliability than OUES, measured using foreshortened data from the 25–75th, 0–50th, 0–70th, and 0–90th percentiles. OUES was similar to OUEP.

No COPD patients showed signs of exercise oscillatory ventilation, so reliability was only assessed in patients with cardiovascular disease. There were 15 patients who fulfilled the criteria for exercise oscillatory ventilation on both tests and a further eight patients who fulfilled the criteria on only one test (ICC 0.72, 95% CI 0.45–0.82).

**Influence of disease aetiology**

The VO2 at AT had a nonsignificant but slightly weaker reliability in CHF over COPD, with a much greater between-test mean difference in the mitral group over the COPD group. There was a significantly improved reliability of the OUES in patients with mitral valve disease; however, ICC was excellent in all groups (≥0.88). The VE/VCO2 ratio at nadir and at AT had better ICCs and smaller mean differences in the COPD and mitral groups compared to the CHF group. Patients with mitral valve disease had a larger mean difference in resting and peak HR between tests, but the ICC was not different between groups, while the HR/VO2 slope was significantly less reliable in the COPD group (Supplementary Table S1).
The effect of ramp protocol

The VE/VCO₂ slope was less reliable in the patients moving to a shallower protocol, with the best ICC in the group moving to steeper protocols. There were differences in the mean difference of DP, circulatory power, maximum watts, and the VO₂/WR slope, the latter showing a decrease in slope in steeper protocols and an increase in slope in shallower protocols. Peak HR and the HR/VO₂ slope displayed significantly stronger ICCs in patients moving to steeper protocols.

The effect of timing

VO₂ at the AT showed a much greater mean difference in patients performing the tests on the same compared
to a different day (mean difference ± SDD, +69 ± 112 vs. −70 ± 120 ml/min). There was almost no mean difference in O₂ pulse when tests were performed on the same day (−0.1 ± 0.8 ml/beat), with a large increase when performed on different days (+0.5 ± 1.0 ml/beat). The VO₂/WR slope was significantly less reliable when measured on different days compared to the same day (ICC 0.46 vs. 0.77, p = 0.02), although mean difference was not significantly different.

Discussion
Cardiopulmonary exercise testing is an established modality in clinical medicine to obtain information on the extent and mechanism of exercise limitation. Good reproducibility of selected variables has been reported. However, numerous variables can be generated for interpretation and the comparative reproducibility of these variables has not been studied. Here we report, using multiple reproducibility measures, the test–retest reliability of a comprehensive set of variables.

Figure 1. Bland–Altman plots
peak VO₂, peak oxygen uptake (ml/min); AT, VO₂ at the anaerobic threshold (ml/min); VE/VCO₂ slope, ventilation/carbon dioxide ratio; OUES, oxygen uptake efficiency slope (l/min per 10-fold increase in ventilation; OUEP, oxygen uptake efficiency plateau; peak O₂ pulse (ml/beat); BR, breathing reserve at peak (%); VO₂/WR slope, oxygen uptake/work rate slope (ml/min/W).

We observed a familiarization effect in a number of variables, with persistently better values on the second test. This effect however, was much smaller than with a previously reported study. Elborn et al. showed a 17% increase in peak VO₂ from test 1 to 2 in heart failure patients, compared with only 2.5% in our study.
The larger difference seen previously may be explained by the use of a treadmill rather than bicycle, where familiarization may be more important. A reproducibility substudy of the HF-ACTION trial did not show a familiarization effect between tests 1 and 2 for patients with CHF in peak VO₂ (vast majority treadmill). However, in the same study other variables, such as exercise time, VE/VCO₂ slope, peak HR, and peak RER did show improvements between tests. Peak VO₂ was similarly unchanged between serial tests in three other studies. Our results agree more closely with these latter studies; while we show a statistically significant familiarization effect for peak VO₂, the magnitude is small and of little clinical relevance. It is interesting to note that the OUES showed no significant difference between the two tests; this variable appeared resilient to familiarization from prior CPX testing. This could be expected given that, as a slope, it is much less effort dependent than other variables such as peak VO₂, or possibly because slopes are typically less affected by ergometer choice. Prior treadmill experience may be more influential for these variables as prior familiarization.

**Different measures of reproducibility**

Most studies on test–retest reproducibility use a single measure, typically the ICC or CoV. The ICC is a ratio of the between-subject variance to the sum of the variance between and within subjects (between tests). ICC will be higher when within-subject variance is low, but will fall when the between-subject variance is small. Hence, two different ICCs can be generated for the same variable by including more diverse participants without a change in the within-subjects’ variance, which is of the greatest interest to clinicians. CPX variables are typically continuous variables with a wide range (e.g. the range of peak VO₂ within our patients was 546.5–2248.9 ml/min).

Despite the limitations of ICC, one benefit, as it is unit-less, is the ability to perform a quantitative comparison of the reproducibility of different variables with different units. However, interpretation must take into account the underlying variance of the compared variables. For example, p_{ET}CO₂ will have low overall variance (which will lower ICC) as an integer value existing over a relatively small range (25–45).

We also present the coefficient of variation. The SDD alone is difficult to interpret because the magnitude of the overall values will heavily influence its relevance. Therefore the SDD/mean ratio gives a measure that is ‘normalized’. However, this normalization process will also inflate a variable’s reproducibility when the range of the variable is an order of magnitude lower than the value, for example arterial oxygen saturations, which will rarely fall below 90% and cannot exceed 100%.

**How the reproducibility of different variables compare**

When measured by ICC, peak VO₂, whether measured as an absolute value, weight adjusted, or percentage of predicted peak, showed excellent test–retest reliability (ICC > 0.91) (Table 2). The O₂ pulse, OUES, OUEP, the VE/VCO₂ ratio at nadir and VCP, breathing reserve, and circulatory power also had excellent ICC values. Despite previous assertions that OUES measured using submaximal data is reliable, it did not compare favourably with full data OUES or peak VO₂. The VO₂ at the AT was not as reliable as many other variables. It has been suggested that the presence of exercise-induced shunting could explain an increased variability in the timing of the AT; however, no exercise-induced shunting was found within our patients to help explain this.

Peak VO₂ also had good test–retest reliability when assessed using CoV. The O₂ pulse also showed good values for CoV, although the OUES was slightly worse. Measurements of the VE/VCO₂ relationship all appeared to have better test–retest reliability via this method (CoV 0.06–0.11), although it can be argued that this is less appropriate than the ICC given the nature of these variables (i.e. they cannot fall close to zero).

The Bland–Altman plots, which allow a more detailed qualitative view of the reproducibility of a variable by plotting the mean of the two tests of each patient against the difference between these tests and the limits of agreement, showed excellent test–retest reliability for peak VO₂ and OUES but with a suggestion of divergence of the residuals as values got larger. However, this effect appeared greater for the VE/VCO₂ slope and ratio at the nadir, while the AT and VO₂/WR relationship showed a greater dispersion around the mean.

**How potential modifying factors can affect reproducibility**

Six potential modifying factors were assessed for their effect on test–retest reliability: age, gender, BMI, disease aetiology, change in ramp protocol, and intertest interval. The test–retest reliability of most variables was resistant to the influences of age, gender, and BMI. Intertest interval did not affect the majority of variables’ ICCs.

The two determinants that made the most difference were disease category and ramp protocol. However, within the disease category, there was no single disease that consistently outperformed the others.
For the ramp protocol, it was largely the patients moving to the steepest ramp that displayed the strongest test–retest reliability. This may be because change in ramp protocol was not randomly chosen; patients with the greatest exercise capacity moved to a more appropriate steeper ramp. This may therefore represent greater reliability amongst patients with better overall exercise capacities.

The VO₂/WR slope was the most affected variable, modified by four of the six influences, in keeping with its overall poor reproducibility.

Study limitations
Patients were aware that test 1 was a familiarization test and may have underperformed on this test. However, the difference in peak VO₂ is only 2.5%, and peak RER was not significantly different between tests, suggesting equivalent effort. Some patients had undergone previous CPX testing (largely treadmill) which may have lessened the impact of familiarization.

Patients were not randomly allocated a test-2 protocol. This was decided, as per study design, based on the first test performance. Hence, both the least and most limited patients would have had a protocol change, while those around median capacity would not. It may be that exercise capacity, rather than protocol change per se, led to some variables dependence on protocol.

The decision to perform both tests on the same day or different days was not randomized, with the subjects deciding if they wished to perform the tests over 1 or 2 days.

Conclusions
The majority of variables calculated from CPX show excellent test–retest reliability. This study estimates, for the first time, multiple values of reproducibility for many variables and allows direct comparison between variables. Peak VO₂, OUES, O₂ pulse, peak circulatory power, and the VE/VCO₂ ratio at nadir show excellent test–retest reliability. Measures of the anaerobic threshold and the VE/VCO₂ slope show good test–retest reliability. The VE/VCO₂ ratio at nadir is superior to the slope, while OUES using foreshortened data is inferior to the full OUES. Peak VO₂ shows a small, albeit significant, familiarization effect, unlikely to be of clinical relevance. OUES does not display this familiarization effect. Overall, there appears to be little necessity for a preliminary test in the majority of patients. Finally, the difference in variables between two tests is largely unaffected by age, gender, BMI, aetiology of disease, different protocol, and intertest interval.

Funding
This work was supported by the British Heart Foundation (FS/04/079 to DPF; PG/11/36/28883 to AB and RW).

Conflict of interest
The authors declare that there is no conflict of interest.

References

