Inter-observer reliability of preoperative cardiopulmonary exercise test interpretation: a cross-sectional study

T. E. F. Abbott,[1,2*] M. Gooneratne,[2*] J. McNeil,[2] A. Lee,[1] D.Z.H Levett,[3] M.P.W. Grocott,[3] M. Swart,[4] and N. MacDonald,[2] for the ARCTIC study investigators[§].

*= joint first authors

§= Members of study group listed at the end of the manuscript

1. *William Harvey Research Institute, Queen Mary University of London, UK*
2. *Barts Health NHS Trust, London, UK*
3. *Critical Care Research Group, Southampton NIHR Biomedical Research Centre, University Hospital Southampton – University of Southampton, Southampton, UK*
4. *South Devon Healthcare NHS Trust, Torbay, UK*

Correspondence to:

Dr Tom Abbott, MRCP

William Harvey Research Institute

Queen Mary University of London

London EC1M 6BQ

e-mail: t.abbott@qmul.ac.uk

Tel: +44 203 594 0352

Keywords: Observational study; Surgery; cardiopulmonary exercise testing.

Main text: 2,857 words **Summary: 245 words**

Summary

Background

Despite the increasing importance of cardiopulmonary exercise testing (CPET) for preoperative risk assessment, the reliability of CPET interpretation is unclear. We aimed to assess the inter-observer reliability of preoperative CPET.

Methods

Prospective, multi-centre, observational study of preoperative CPET interpretation. Participants were professionals with previous experience or training in CPET, assessed by a standardised questionnaire. Each participant interpreted 100 CPETs using standardised software. The CPET variables of interest were oxygen consumption at the anaerobic threshold (AT) and peak oxygen consumption (VO_2peak). Inter-observer reliability was measured using intra-class correlation coefficient (ICC) with a random effects model. Results were presented as ICC with 95% confidence interval, where ICC of 1 represents perfect agreement and ICC of 0 represents no agreement.

Results

8/28 (28.6%) participants were clinical physiologists, 10 (35.7%) were junior doctors and 10 (35.7%) were consultant doctors. The median previous experience was 140 (IQR 55-700) CPETs. After excluding the first 10 CPETs (acclimatisation) for each participant and missing data, the primary analysis of AT and VO_2peak included 2125 and 2414 CPETS respectively. Inter-observer agreement for numerical values of AT (ICC 0.83 [0.75-0.90]) and VO_2peak (ICC 0.88 [0.84-0.92] was good. In post-hoc analysis, inter-observer agreement for identification of the presence of a reportable AT was excellent (ICC 0.93 [0.91-0.95]) and a reportable VO_2peak was moderate (0.73 [0.64-0.80]).

Conclusions

Inter-observer reliability of interpretation of numerical values of two commonly used CPET variables was good (>80%). However, inter-observer agreement regarding the presence of a reportable value was less consistent.

Introduction

More than 1.5 million major surgical procedures are carried out in the United Kingdom (UK) every year.[1, 2] Estimates of postoperative mortality range from 1 to 4% depending on the population sampled and the type of surgical procedure.[1, 3, 4] However, it is clear that mortality and morbidity following surgery are greater in high-risk cohorts, where patients have pre-existing medical conditions, are elderly or undergoing a major surgical procedure, for example surgery to gastrointestinal tract.[5, 6] Postoperative morbidity is associated with reduced long-term survival and is likely to have a lasting impact of subsequent quality of life.[7, 8]

In the UK, cardiopulmonary exercise testing (CPET) is increasingly used for risk assessment before major surgery.[9-13] The majority of preoperative CPET clinics use protocols based on consensus guidelines.[14] However, while preoperative exercise capacity has been associated with morbidity and mortality after major surgery, it remains unclear which CPET-derived variable is best for predicting outcome after major surgery.[15, 16] Two of the most commonly used variables are the anaerobic threshold (AT), an index of sustainable, submaximal exercise capacity, and peak oxygen consumption (VO_2 peak), an index of maximal exercise capacity.[17] The AT is the point during an incremental exercise test above which arterial lactate rises in a sustained manner above resting levels,[18] while VO_2 peak is the highest oxygen uptake attained at end-exercise.[19] Both can be estimated non-invasively using respiratory gas analysis.[19-21] The reliability of CPET interpretation, particularly the AT, has been questioned;[22] but, this has been subject to only limited investigation. Few studies have investigated inter-observer error associated with CPET,[23] with only one specifically focused on CPET before surgery.[24] These studies were limited to the AT, and so did not report the reliability of VO_2 peak measurement, nor did they take into account the experience or training of the observers.

Despite the increasing importance placed on CPET for preoperative risk assessment, there is little evidence of reliability of interpretation between clinicians. Variations in the reported values of CPET variables could exert significant influence on perioperative care planning.[25, 26] Therefore, in this prospective study, we investigated inter-observer reliability of anaerobic threshold and peak oxygen consumption identification, and the relative influence of training and experience in CPET interpretation.

Methods

Study design and setting

This was a prospective, multi-centre, observational study of inter-observer reliability of preoperative cardiopulmonary exercise test (CPET) interpretation, where inter-observer reliability refers to the consistency of agreement between observers. The study received research ethics approval (QMREC1531a) and was conducted in accordance with the principles of the Declaration of Helsinki and the Research Governance Framework. Reporting is consistent with the STROBE and STARD guidelines for observational studies and studies of diagnostic accuracy.[27, 28]

Participants (observers)

Observers were professionals with previous experience or training conducting or reporting CPETs. Observers were identified and recruited, by approaching UK hospitals known to have a preoperative CPET service, through professional networks, and by word of mouth. All observers gave written informed consent before taking part in the study.

Study conduct and data collection

Each observer interpreted the oxygen consumption at the anaerobic threshold and the peak oxygen consumption, using the electronic records of 100 previously conducted preoperative CPETs from a dedicated research database. CPET data were viewed using ZAN software (NSpire Health, UK). All observers were given the same set of generic instructions and were asked to interpret the CPET data using the method(s) they would ordinarily use. The ZAN software allowed the following methods for assessment of AT to be used: V-slope, modified V-slope, ventilatory equivalents, excess carbon dioxide, or respiratory exchange ratio. The default settings for the ZAN software used 30-second data averaging, although each participant was able to change this. Every observer viewed the first ten CPETs in the same order, starting with

test number 1 and finishing with test number 10. This acted as acclimatization to the software and environment. Thereafter, each observer viewed the subsequent 90 CPETs in a random order unique to each participant. Observers recorded their results directly into an Excel pro-forma (Microsoft, Redmond, USA) and completed a short questionnaire about their previous training and experience with CPET interpretation (supplementary table 1). During the testing a member of the research team was available in order to respond to problems, for example software malfunctions.

CPET database generation

Each observer interpreted the same electronic research database containing raw data from 100 CPETs. These were chosen at random from a preoperative assessment clinic database consisting of ~250 cases. CPETs were briefly screened for data completeness and plausibility prior to inclusion by study investigators (TA, MG), but were not fully interpreted to minimise investigator bias. Upon entry into the research database, each CPET record was fully anonymised, including the removal of patient identifiable data, and assigned a unique study ID number.

Key variables

The variables of interest were oxygen consumption at the anaerobic threshold and peak oxygen consumption, both measured in $ml.kg.min^{-1}$ and identified using the method(s) each observer would ordinarily use in their clinical practice.

Statistical analysis

The analysis was prospectively planned before the data were reviewed. We used Python [www.python.org] to compile a results database and STATA version 14 (STATACorp LP, Texas, USA) to analyse the data. The first 10 CPETs used to calibrate each observer were not included in the primary analysis. We used the intra-class correlation coefficient (ICC) with a two-way

random effects model for absolute agreement to measure inter-observer reliability; this accounted for the fact that our sample of observers was derived from a larger population of professionals who interpret CPET. We reported the average absolute ICC across the whole group of observers. Firstly, we calculated ICC for the whole sample. Secondly, we stratified the sample according to the following measures of expertise in CPET interpretation: self-rating (novice, inexperienced, experienced, very experienced, expert), total number of tests interpreted (≤55, 56-140, 141-700, >700), years of experience interpreting CPETs (≤1, 2-3, 4-5, ≥6) and profession (physiologist, consultant doctor, junior doctor). Thresholds for continuous data were defined by dividing the data into quartiles and were not arbitrarily defined *a priori*. We calculated the ICC for each strata. Results were presented as intra-class correlation coefficient (ICC), where 0 indicates no agreement and 1 indicates perfect agreement, with 95% confidence intervals. ICC values were interpreted according to the classification of reliability by Koo et al: <0.50, poor; 0.50-0.75, moderate; 0.75-0.90, good; and >0.90, excellent. Normally distributed data were expressed as mean +/- standard deviation (SD) and non-normally distributed data were expressed as median +/- interquartile range (IQR). Binary data were expressed as percentages.

Sensitivity analyses

We repeated the analysis after including the first 10 CPETs that were excluded from the primary analysis. We repeated the primary analysis for the following additional measures of experience: number of CPETs interpreted in the last year (≤28, 29-80, 81-150, ≥150), number of CPETs interpreted per week (≤2, 3, 4, ≥5) and attendance at a formal CPET training course.

Statistical power

Guidelines for using intra-class correlation to measure reliability suggest obtaining 30 individual measurements from at least three observers.[29] Assuming a type 1 error rate of 5%, our sample of 28 observers, each interpreting 100 cases, gives >99% power to identify excellent agreement

(ICC 0.90-0.99), 89% power to detect good agreement (ICC 0.75-0.90) and 95% power to detect moderate agreement (ICC 0.50-0.75). Power calculations used STATA *sampicc* function.

Results

Twenty-eight observers were recruited into the study between 17th August 2015 and 13th March 2017. The primary analysis included 2,125 observations for oxygen consumption at the anaerobic threshold and 2,414 observations for peak oxygen consumption, after excluding the first 10 CPETs for each observer and any missing data (figure 1). Anaerobic threshold data was not recorded for 395/2520 tests (16%) and peak oxygen consumption was not recorded for 106/2520 teats (4%). The baseline characteristics of the cohort are described in table 1. 8/28 (28.6%) observers were physiologists, 10/28 (35.7%) were junior (non-consultant) doctors and 10/28 (35.7%) were consultant doctors, from 16 institutions. The median (IQR) duration of previous experience interpreting CPETs was 2.5 (1-5) years; 3/28 (10.7%) considered themselves experts, while 4/28 (14.3%) considered themselves novices.

Across the whole cohort, considering all interpreted values for all patients, the mean and median of anaerobic threshold were 10.9 (SD. 2.9) ml/kg/min and 10.6 (IQR. 9.0-12.4) ml/kg/min respectively, while the mean and median of peak oxygen consumption were 14.7 (SD. 3.9) ml/kg/min and 14.0 (IQR. 12.3-16.9) respectively. The average number of valid observations per case was 21 for anaerobic threshold and 24 for peak oxygen consumption. The median interpreted values and ranges of anaerobic threshold and peak oxygen consumption for each patient is shown in figure 2 and stratified by years of experience in supplementary figures 1 and 2. Across the whole cohort, the ICC for oxygen consumption at the anaerobic threshold was 0.83 [0.75-0.90] and 0.88 [0.84-0.92] for peak oxygen consumption, indicating good reliability between observers, according to the scale described by Koo et al.[29] Table 2 shows the ICC stratified by level of experience with CPET interpretation according to several different categorisations. For example, observers that considered themselves expert had an ICC of 0.96 [0.93-0.97] for VO_2 peak compared to 0.74 [0.65-0.71] for observers that considered themselves

novice (figure 3). Whereas for the AT, the ICC for expert observers was 0.74 [0.59-0.84] compared to 0.85 [0.79-0.89] for novice observers (p=0.09).

Sensitivity analyses

When we repeated the primary analysis including the first 10 CPETs for each observer that had been removed from the primary analysis, our results were very similar: the ICC for VO_2 at the anaerobic threshold was 0.83 [0.75-0.89] and the ICC for VO_2 peak was 0.90 [0.86-0.93]. We present the results of the primary analysis stratified by the number of CPETs interpreted in the last year, the average number per week and attendance at a formal course in the supplement (supplementary table 2). Out of 2,520 total observations, anaerobic threshold was not reported in 295 cases and peak oxygen consumption was not reported in 106 cases. To examine the degree of inter-observer agreement/disagreement regarding the presence or absence of a reportable anaerobic threshold or peak oxygen consumption value, we undertook a post-hoc analysis. We categorised each observation as either reported or not reported and used the intra-class correlation coefficient (ICC) with a two-way random effects model for consistency of agreement to measure inter-observer reliability. For anaerobic threshold, average agreement between observers was excellent (ICC 0.93 [0.91-0.95]) and for peak oxygen consumption average agreement was moderate (0.73 [0.64-0.80]).

Discussion

The principal finding of this study was that the inter-observer reliability of reporting numerical values of two commonly derived preoperative cardiopulmonary exercise test variables was good.[29] Interpretation of peak oxygen consumption appeared consistent with anaerobic threshold (AT), returning ICCs of 0.89 and 0.83 respectively. However, there was greater heterogeneity of agreement when identifying whether or not a reportable value existed, with excellent agreement for identifying a reportable anaerobic threshold and moderate agreement for identifying a reportable peak oxygen consumption. This suggests that observers are able to identify whether or not there is an AT, but are less consistent at reporting the specific numerical value. In contrast, there appeared to be less agreement when identifying the presence of a reportable VO_2peak, but where this was certain, greater agreement regarding the specific value. The reliability of interpreting numerical values seemed to increase with the experience of the observer, particularly for VO_2peak, although this only appears to be statistically significant when comparing self-rated experts or those with >5 years experience with novices (table 2), while the range of reported values appeared to narrow for observers with >5 years experience (supplementary figure 2). We did not see consistent increases in reliability across all domains of experience, or consistent increases in reliability with increasing experience for the AT. This may be due to heterogeneity of training or variation in the methods used to interpret the AT among observers, with more experienced observers using a wider variety of techniques.[30] Reassuringly, reliability was good even in practitioners with relatively little experience. The reliability of CPET interpretation is similar to echocardiography, where the inter-observer reliability is reported between 80% and 94%.[31-33]

Inter-observer reliability of AT interpretation has been subject to only limited investigation.[23, 24, 34, 35] A study of 1679 patients with heart failure and a smaller study of ten patients undergoing surgery reported agreement of 80-90%, suggesting that the interpretation of AT can be reliable

with experienced practitioners.[2324] However, the later study in surgical patients was limited by the small number of observations made by each observer. Our study examines the reliability of preoperative CPET interpretation in a large number of observations (>2000). Our results suggest that numerical values of anaerobic threshold and peak oxygen consumption can be interpreted with a high degree of reliability by a heterogeneous group of observers with a variety of training and experience. Nine out of ten participants had attended formal training in CPET, which may have contributed to the consistency of interpretation. Our data suggest that experience may influence the reliability of CPET interpretation; a finding that may have implications for clinical decision-making, where small variations in CPET results could change the course of perioperative care. While this study aimed to assess the influence of inter-observer reliability on CPET results, this is not the only source of between-test variation. Other sources include: measurement error due to either equipment or software, different CPET protocols (e.g. ramp selection) and/or physiological variation for any given subject.

Our study has several strengths. This was the largest study of preoperative cardiopulmonary exercise test reliability, of which we are aware. We included observers from a variety of professional backgrounds, a large number of hospitals and with varying levels of experience, making our results generalisible to a large number of professionals that interpret CPET in the UK. For the first time we investigated the potential influence of training and experience on CPET interpretation.

There also are several limitations to our approach. Firstly, all observers interpreted the CPETs using Zan software, which they may not have been familiar with before taking part in the study. To reduce the risk of observer bias, each participant was given a standardised introduction to the study and the software. We further reduced the risk of bias by removing the first ten CPETs, which acted as acclimatisation to the software, from the primary analysis. It is also possible that

the way that the ZAN software handled or presented the CPET data may have influenced our estimation of inter-observer reliability. Since we did not make comparisons with other software, this is difficult to assess. Further research is needed to determine the influence of software differences on CPET results. Secondly, we were vigilant to a potential learning effect, where agreement may have increased observers progressed through the study. To mitigate against this, after the first ten CPETs, each observer interpreted the CPETs in a random order. However, due to the random order, we are unable to perform post-hoc tests of this potential bias. When we compare the results of the primary analysis with and without the first 10 CPETs, the results are similar. Thirdly, we asked observers to complete a questionnaire regarding their previous training or experience in CPET interpretation. Since these self-reported data are not contemporaneous, they may be inaccurate or subject to recall bias, which may influence the results of our secondary analysis. The vast majority (89%) of participants indicated that they had undergone previous formal training in CPET interpretation. However, the heterogeneity of training and the influence on our results is unknown. Further research could be directed at the influence of specific types of training on the interpretation of CPET data. Where measures of experience were continuous data, for example the total number of CPETs interpreted, we identified cut-points by dividing the cohort into quartiles, rather than using *a priori* thresholds. Fourthly, to reduce bias when setting up the CPET database, cases were selected using a randomised process. In addition, investigators briefly screened the raw patient data before inclusion, however to minimise observer bias, they did not fully interpret the data before the start of the study. It is possible that some CPETs in the research database did not have a measurable anaerobic threshold or peak oxygen consumption. Our results, which agree with previously published data, indicating that peak oxygen consumption was reported more often than the anaerobic threshold, suggesting that there may be more tests where the anaerobic threshold could not be identified compared to peak oxygen consumption. This may account for the small amount of missing data in the sample.[36] Finally, for pragmatic reasons, the analysis

was restricted to two commonly used CPET variables: anaerobic threshold and peak oxygen consumption. However, we recognise that there are many other CPET variables used in clinical practice, which might be more or less at risk of observer bias than the ones we tested. Further research would be needed to evaluate the influence of inter-observer variation on the interpretation of any additional CPET variables.

Conclusion

Interpretation of numerical values of two commonly used CPET variables has good (>80%) inter-observer reliability. However, inter-observer agreement regarding whether that value could be reported was less consistent. The reliability of interpreting numerical values of VO_2 peak may be influenced by the experience of observers, although this was not consistent across all domains. Reliability of AT interpretation did not appear to vary with the experience of observers.

Conflict of interest statement

TEFA is a committee member of the Perioperative Exercise Testing and Training Society. All other authors declare no conflicts of interest.

Authors' contributions

TEFA and JM conceived the study. TEFA and NM designed the protocol, with advice from MPG, DL and MS. TEFA, MG, NM and MS collected the data. TEFA analysed the data with input from AL and NM. The manuscript was drafted by TEFA, MG and NM, and revised following critical review by all authors.

Sources of funding

TEFA was supported by a Medical Research Council and British Journal of Anaesthesia clinical research training fellowship. AL was supported by the NIHR Biomedical Research Centre at Barts Health NHS Trust and a "SmartHeart" EPSRC programme grant (EP/P001009/1).

Acknowledgements

The authors would like to thank NSpire Health (Hertford, UK) for providing a free copy of the ZAN software to facilitate this research project.

ARCTIC study group

Tom Abbott, Neil MacDonald, Mevan Gooneratne, Ashok Raj, Martin Rooms, Maggie Nicol, Maria Koutra, Stephen Halworth, Kim Wilkins, Daniel Nevin, Eleanor Gaultry, James Otto, Dan Bell, Marie Hardy, Pradeep Prabhu, Rao Ravishankar, Joao Correia, Christian Beilstein, Kathryn Greaves, Stephen James, Hannah Tighe, Joe Perks, James Pennington, Andy Pritchard, Peter Moxon, Katherine Brown, Michael Swart, John Carlisle.

References

1 Abbott TEF, Fowler AJ, Dobbs T, et al. Frequency of surgical treatment and related hospital procedures in the United Kingdom: A national ecological study using hospital episode statistics. *British journal of anaesthesia* 2017

2 Weiser TG, Haynes AB, Molina G, et al. Estimate of the global volume of surgery in 2012: an assessment supporting improved health outcomes. *Lancet* 2015; **385 Suppl 2**: S11

3 Pearse RM, Moreno RP, Bauer P, et al. Mortality after surgery in Europe: a 7 day cohort study. *Lancet* 2012; **380**: 1059-65

4 International Surgical Outcomes Study g. Global patient outcomes after elective surgery: prospective cohort study in 27 low-, middle- and high-income countries. *British journal of anaesthesia* 2016; **117**: 601-9

5 Pearse RM, Harrison DA, James P, et al. Identification and characterisation of the high-risk surgical population in the United Kingdom. *Critical care* 2006; **10**: R81

6 Grocott MP, Pearse RM. Perioperative medicine: the future of anaesthesia? *British journal of anaesthesia* 2012; **108**: 723-6

7 Head J, Ferrie JE, Alexanderson K, et al. Diagnosis-specific sickness absence as a predictor of mortality: the Whitehall II prospective cohort study. *Bmj* 2008; **337**: a1469

8 Khuri SF, Henderson WG, DePalma RG, et al. Determinants of long-term survival after major surgery and the adverse effect of postoperative complications. *Annals of surgery* 2005; **242**: 326-41; discussion 41-3

9 Minto G, Struthers RA. It's not about the bike: enhancing oxygen delivery. *British journal of anaesthesia* 2017; **118**: 655-7

10 Hennis PJ, Meale PM, Grocott MP. Cardiopulmonary exercise testing for the evaluation of perioperative risk in non-cardiopulmonary surgery. *Postgraduate medical journal* 2011; **87**: 550-7

11 James S, Jhanji S, Smith A, O'Brien G, Fitzgibbon M, Pearse RM. Comparison of the prognostic accuracy of scoring systems, cardiopulmonary exercise testing, and plasma biomarkers: a single-centre observational pilot study. *British journal of anaesthesia* 2014; **112**: 491-7

12 Older P, Hall A, Hader R. Cardiopulmonary exercise testing as a screening test for perioperative management of major surgery in the elderly. *Chest* 1999; **116**: 355-62

13 West MA, Asher R, Browning M, et al. Validation of preoperative cardiopulmonary exercise testing-derived variables to predict in-hospital morbidity after major colorectal surgery. *The British journal of surgery* 2016; **103**: 744-52

14 American Thoracic S, American College of Chest P. ATS/ACCP Statement on cardiopulmonary exercise testing. *American journal of respiratory and critical care medicine* 2003; **167**: 211-77

15 Wijeysundera DN, Pearse RM, Shulman MA, et al. Measurement of Exercise Tolerance before Surgery (METS) study: a protocol for an international multicentre prospective cohort study of cardiopulmonary exercise testing prior to major non-cardiac surgery. *BMJ open* 2016; **6**: e010359

16 Abbott TEF, Minto G, Lee AM, Pearse RM, Ackland G. Elevated preoperative heart rate is associated with cardiopulmonary and autonomic impairment in high-risk surgical patients. *British journal of anaesthesia* 2017; **119**: 87-94

17 Otto JM, Plumb JOM, Wakeham D, et al. Total haemoglobin mass, but not haemoglobin concentration, is associated with preoperative cardiopulmonary exercise testing-derived oxygen-consumption variables. *British journal of anaesthesia* 2017; **118**: 747-54

18 Wasserman K, McIlroy MB. Detecting the Threshold of Anaerobic Metabolism in Cardiac Patients during Exercise. *The American journal of cardiology* 1964; **14**: 844-52

19 Day JR, Rossiter HB, Coats EM, Skasick A, Whipp BJ. The maximally attainable VO2 during exercise in humans: the peak vs. maximum issue. *Journal of applied physiology* 2003; **95**: 1901-7

20 Beaver WL, Wasserman K, Whipp BJ. A new method for detecting anaerobic threshold by gas exchange. *Journal of applied physiology* 1986; **60**: 2020-7

21 Older PO, Levett DZH. Cardiopulmonary Exercise Testing and Surgery. *Annals of the American Thoracic Society* 2017; **14**: S74-S83
22 Hopker JG, Jobson SA, Pandit JJ. Controversies in the physiological basis of the 'anaerobic threshold' and their implications for clinical cardiopulmonary exercise testing. *Anaesthesia* 2011; **66**: 111-23
23 Myers J, Goldsmith RL, Keteyian SJ, et al. The ventilatory anaerobic threshold in heart failure: a multicenter evaluation of reliability. *Journal of cardiac failure* 2010; **16**: 76-83
24 Sinclair RC, Danjoux GR, Goodridge V, Batterham AM. Determination of the anaerobic threshold in the pre-operative assessment clinic: inter-observer measurement error. *Anaesthesia* 2009; **64**: 1192-5
25 Levett DZ, Grocott MP. Cardiopulmonary exercise testing, prehabilitation, and Enhanced Recovery After Surgery (ERAS). *Canadian journal of anaesthesia = Journal canadien d'anesthesie* 2015; **62**: 131-42
26 Swart M, Carlisle JB. Case-controlled study of critical care or surgical ward care after elective open colorectal surgery. *The British journal of surgery* 2012; **99**: 295-9
27 Bossuyt PM, Reitsma JB, Bruns DE, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *Bmj* 2015; **351**: h5527
28 von Elm E, Altman DG, Egger M, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Bmj* 2007; **335**: 806-8
29 Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of chiropractic medicine* 2016; **15**: 155-63
30 Dickstein K, Barvik S, Aarsland T, Snapinn S, Karlsson J. A comparison of methodologies in detection of the anaerobic threshold. *Circulation* 1990; **81**: II38-46
31 Herold IH, Saporito S, Bouwman RA, et al. Reliability, repeatability, and reproducibility of pulmonary transit time assessment by contrast enhanced echocardiography. *Cardiovascular ultrasound* 2016; **14**: 1
32 Saul T, Avitabile NC, Berkowitz R, et al. The Inter-rater Reliability of Echocardiographic Diastolic Function Evaluation Among Emergency Physician Sonographers. *The Journal of emergency medicine* 2016; **51**: 411-7
33 Carlton EF, Sontag MK, Younoszai A, et al. Reliability of Echocardiographic Indicators of Pulmonary Vascular Disease in Preterm Infants at Risk for Bronchopulmonary Dysplasia. *The Journal of pediatrics* 2017
34 Gladden LB, Yates JW, Stremel RW, Stamford BA. Gas exchange and lactate anaerobic thresholds: inter- and intraevaluator agreement. *Journal of applied physiology* 1985; **58**: 2082-9
35 Matsumura N, Nishijima H, Kojima S, Hashimoto F, Minami M, Yasuda H. Determination of anaerobic threshold for assessment of functional state in patients with chronic heart failure. *Circulation* 1983; **68**: 360-7
36 Lai CW, Minto G, Challand CP, et al. Patients' inability to perform a preoperative cardiopulmonary exercise test or demonstrate an anaerobic threshold is associated with inferior outcomes after major colorectal surgery. *British journal of anaesthesia* 2013; **111**: 607-11

Table 1. Characteristics of observers.

Continuous data expressed as median with interquartile range (IQR), categorical data expressed as frequencies with percentages. CPET, cardiopulmonary exercise test.

Years experience, median (IQR)	2.5 (1-5)
CPETs interpreted (total), median (IQR)	140 (55-700)
CPETs interpreted (per year), median (IQR)	80 (28-150)
CPETs interpreted (per week), median (IQR)	3 (2-4)
Attendance at formal CPET course	25 (89%)
Self rated experience	
Novice	4 (14.3%)
Inexperienced	9 (32.1%)
Experienced	10 (35.7%)
Very experienced	2 (7.1%)
Expert	3 (10.7%)
Profession/grade	
Physiologist	8 (28.6%)
Junior doctor	10 (35.7%)
Consultant doctor	10 (35.7%)

Table 2. Reliability of CPET interpretation.

Intra-class correlation coefficient (ICC), with 95% confidence intervals (95% CI), for oxygen consumption at the anaerobic threshold (VO_2 at AT) and peak oxygen consumption (VO_2 peak). Results presented for the whole cohort and stratified by different measures of experience interpreting cardiopulmonary exercise tests (CPETs).

	VO_2 at AT ICC (95% CI)	VO_2 peak ICC (95% CI)
Whole cohort	0.83 (0.75-0.90)	0.88 (0.84-0.92)
Years experience (quartiles)		
≤1	0.82 (0.74-0.88)	0.84 (0.79-0.88)
2-3	0.78 (0.69-0.86)	0.91 (0.87-0.94)
4-5	0.78 (0.64-0.87)	0.87 (0.72-0.93)
≥6	0.78 (0.67-0.86)	0.95 (0.92-0.97)
CPETs interpreted in total (quartiles)		
≤55	0.82 (0.74-0.88)	0.98 (0.97-0.99)
56-140	0.84 (0.78-0.89)	0.90 (0.85-0.93)
141-700	0.76 (0.65-0.85)	0.82 (0.75-0.88)
>700	0.78 (0.67-0.86)	0.94 (0.91-0.96)
Self rated experience		
Novice	0.85 (0.79-0.89)	0.74 (0.65-0.81)
Inexperienced	0.82 (0.73-0.89)	0.88 (0.84-0.92)
Experienced	0.75 (0.67-0.83)	0.93 (0.89-0.96)
Very experienced	0.86 (0.77-0.91)	0.94 (0.89-0.96)
Expert	0.74 (0.59-0.84)	0.96 (0.93-0.97)
Profession/grade		
Physiologist	0.82 (0.74-0.89)	0.83 (0.76-0.89)
Junior doctor	0.82 (0.75-0.88)	0.87 (0.85-0.92)
Consultant doctor	0.70 (0.59-0.80)	0.96 (0.94-0.97)

Figure legends

Figure 1. Flow diagram showing the number of cardiopulmonary exercise tests (CPETs) included in each analysis. VO2 peak = peak oxygen consumption.

Figure 2. Scatter plot for median anaerobic threshold (upper panel) and median peak oxygen consumption (lower panel) in ml/kg/min. Each dot represents a single patient. The cohort is ordered according to increasing values of the median, with error bars indicating the range of values. Median of anaerobic threshold for the whole cohort was 10.6, with range of medians 3.0 - 20.4 ml/kg/min. The median peak oxygen consumption for the whole cohort was 14.0 with range of medians 4.0 – 29.7 ml/kg/min.

Figure 3. Bar chart showing intra-class correlation coefficient (ICC) for agreement between observers, stratified by self-reported experience of CPET interpretation. Peak oxygen consumption (blue bars) and anaerobic threshold (green bars). An ICC of 0.00 represents no agreement and an ICC of 1.00 represents perfect agreement.

Figure 4. Bar chart showing intra-class correlation coefficient (ICC) for agreement between observers, stratified by quartiles of self-reported total number of CPETs conducted by each observer. Peak oxygen consumption (blue bars) and anaerobic threshold (green bars). An ICC of 0.00 represents no agreement and an ICC of 1.00 represents perfect agreement.