BJA
British Journal of Anaesthesia

The cardiopulmonary exercise test grey zone; optimising fitness stratification by application of critical difference

Journal:	*British Journal of Anaesthesia*
Manuscript ID	BJA-2017-00683-JT064.R2
Article Type:	Clinical Investigation
Date Submitted by the Author:	08-Feb-2018
Complete List of Authors:	Rose, George; University of South Wales, Faculty of Life Sciences and Education Davies, Richard; Cardiff and Vale University Health Board, Department of Anaesthetics Davison, Gareth; Ulster University, Faculty of Life and Health Sciences Adams, Richard; Cardiff University, Department of Medicine; Velindre Cancer Centre, Oncology Williams, Ian; Cardiff and Vale University Health Board, Department of Surgery Lewis, Michael; University of South Wales, Faculty of Life Sciences and Education; Cwm Taf University Health Board, Department of Surgery Appadurai, Ian R.; Cardiff and Vale University Health Board, Department of Anaesthetics Bailey, Damian; University of South Wales, Faculty of Life Sciences and Education
Mesh keywords:	Anaerobic threshold, cardiopulmonary exercise test, Risk assessment

The cardiopulmonary exercise test grey zone; optimising fitness stratification by application of critical difference

G. A. Rose[1*], R. G. Davies[2], G. W. Davison[3], R. A. Adams[4], I. M. Williams[5], M. H. Lewis[6], I. R. Appadurai[2], and D. M. Bailey[1*]

[1]Neurovascular Research Laboratory, Faculty of Life Sciences and Education, University of South Wales, UK; [2]Department of Anaesthetics, University Hospital of Wales, Cardiff, UK; [3]Sport and Exercise Sciences Research Institute, Ulster University, Newtownabbey, NI, UK; [4]School of Medicine, Cardiff University, and Velindre Cancer Centre, Cardiff, UK; [5]Department of Surgery, University Hospital of Wales, Cardiff, UK; [6]Department of Surgery, Royal Glamorgan Hospital, Llantrisant, UK.

Running title: Cardiorespiratory fitness and surgical risk stratification

***Correspondence:**

Neurovascular Research Laboratory, Faculty of Life Sciences and Education, University of South Wales, United Kingdom, CF37 4AT. Tel: +44 (0)1443-652296; Fax: +44 (0)1443-652285; *e*mail: damian.bailey@southwales.ac.uk or george.rose@southwales.ac.uk

Abstract

Background. Cardiorespiratory fitness (CRF) can inform patient care, though to what extent natural variation in CRF influences clinical practice remains to be established. We calculated natural variation for cardiopulmonary exercise test (CPET) metrics, which may have implications for fitness stratification.

Methods. In a two-armed experiment, critical difference (CD) comprising analytical imprecision and biological variation was calculated for CRF, and thus defined the magnitude of change required to claim a clinically meaningful change. This metric was retrospectively applied to 213 patients scheduled for colorectal surgery. These patients underwent CPET and the potential for misclassification of fitness was calculated. We created a model with boundaries inclusive of natural variation (CD applied to oxygen uptake at anaerobic threshold ($\dot{V}O_2$-AT): 11mL O_2 kg^{-1} min^{-1}, peak oxygen uptake ($\dot{V}O_2$ peak): 16mL O_2 kg^{-1} min^{-1}, and ventilatory equivalent for carbon dioxide at AT ($\dot{V}_E/\dot{V}CO_2$-AT): 36).

Results. The CD for $\dot{V}O_2$-AT, $\dot{V}O_2$ peak, and $\dot{V}_E/\dot{V}CO_2$-AT was 19%, 13%, and 10%, resulting in false negative and false positive rates of up to 28 and 32% for unfit patients. Our model identified boundaries for unfit and fit patients: AT < 9.2 and ≥ 13.6mL O_2 kg^{-1} min^{-1}, $\dot{V}O_2$ peak < 14.2 and ≥ 18.3mL kg^{-1} min^{-1}, $\dot{V}_E/\dot{V}CO_2$-AT ≥ 40.1 and < 32.7, between which an area of indeterminate-fitness was established. With natural variation considered, up to 60% of patients presented with indeterminate-fitness.

Conclusions. These findings support a reappraisal of current clinical interpretation of CRF highlighting the potential for incorrect fitness stratification when natural variation is not accounted for.

Key words: Anaerobic threshold; cardiopulmonary exercise test; risk assessment.

Introduction

Cardiopulmonary exercise testing (CPET) is a non-invasive procedure to determine the level of cardiorespiratory fitness (CRF) of patients during a progressive exercise challenge to symptom limited maximum. CPET is used as a tool for preoperative assessment of physical fitness for intra-abdominal surgery to aid clinical decision-making given its increasingly proven association with post-operative outcome.[1-7] Furthermore, The American Heart Association recently published a scientific statement promoting cardiorespiratory fitness (CRF) as a clinical vital sign.[8] Despite increasing support for CPET, the mechanisms underpinning CRF that provide protection require further investigation.

The seminal work of Older and colleagues documented an 18% mortality rate in elderly surgical patients with a pulmonary oxygen uptake at the anaerobic threshold ($\dot{V}O_2$-AT) of < 11mL oxygen (O_2) kg^{-1} (total body mass) min^{-1} compared to 0.8% recorded in patients with a $\dot{V}O_2$-AT ≥ 11mL O_2 kg^{-1} min^{-1}.[9] Other biomarkers including peak oxygen uptake ($\dot{V}O_2$ peak) <15mL O_2 kg^{-1} min^{-1} and ventilatory equivalent for carbon dioxide at AT ($\dot{V}_E/\dot{V}CO_2$-AT) > 42 have predicted post-operative survival following abdominal aortic aneurysm surgery.[2] Studies have further attempted to define threshold values in an effort to optimise risk prediction; for example a range of AT values from 9.0 to 11mL O_2 kg^{-1} min^{-1} have been reported, [4 5 9-12] thus demonstrating that variation is present and that a single cut-point cannot be recommended.

Like most biomarkers, CRF is a dynamic metric subject to natural variation and thus needs to be interpreted with caution. Such variation encompasses both analytical and biological components that collectively contribute to the critical difference (CD) as originally described by Fraser and Fogarty.[13] The CD represents random variation around a homeostatic point indicative of the change that must occur before a true difference of clinical significance can be claimed. The concept of CD, yet to be applied to clinical CPET variables, emanates

from the field of clinical biochemistry and has been applied to metabolic biomarkers of exercise stress and clinical patients.[14 15]

The current study reflects the first attempt within the clinical setting to quantify the CD of established CPET markers of CRF with corresponding implications for patient management. We hypothesise that natural variation is present in markers of CRF, and will thus impact upon patient fitness stratification.

Methods

Ethical approval

The University of South Wales Ethics Committee (LSE1636GREO), and Cardiff and Vale University Health Board (15/AIC/6352) approved the study. All procedures were carried out in accordance with the Declaration of Helsinki of the World Medical Association.[16] Written informed consent was obtained from participants in study arm 1. Study arm 2 constituted a retrospective analysis of an anonymized database and thus patient consent was waived.

Design

We conducted a two-armed study. First, in order to determine the CDs of selected CPET variables (reported as independent predictors of post-operative outcome), analytical variation was calculated, and biological variation derived using repeated CPET results from a young apparently healthy population (Arm 1). Subsequently, these CD values were retrospectively applied to an anonymised database of patients who had CPET prior to colorectal surgery, in order to re-appraise fitness stratification (Arm 2).

Study arm 1: CD determination

Analytical variation (CV_A); the first component of CD, was determined by repeatedly passing inspired and expired gases through a Medgraphics Ultima metabolic cart (MedGraphics™, Gloucester, UK) in a manner that replicated typical ventilatory responses during the latter stages of a patient CPET (ie. pulmonary minute ventilation of 25 $L.min^{-1}$). In a series of eight repeated trials each lasting ten respiratory cycles, a 250 L Douglas bag containing saturated expired gas (17% O_2, 5% CO_2) and an equivalent volume of ambient gas was passed through a pneumotach and gas analyser. Inspiration and expiration were simulated using two-way non-rebreathing valves (2700 Series) connected to two factory calibrated 3 L syringes (Hans

Rudolph, Kansas City, USA) operated simultaneously (Figure 1.). Prior to sampling, calibration was undertaken in accordance with manufacturer's guidelines using a 3 L syringe and a known precision gas. During data collection the middle five of seven breaths were averaged.

The within participant coefficient of variation (CV_W) from which biological variation could be calculated, was determined by completion of three repeat CPETs separated by a minimum of 24 hours, for 12 healthy participants (Table 1). Tests were conducted in a randomised order at three time points across operating hours for patient CPET clinics (09:00 to 10:30, 12:00 to 13:30, and 15:00 to 17:00). All CPETs were conducted to volitional fatigue using the Wasserman protocol,[17] the same metabolic cart and investigator, and calibration undertaken as previously described. Following three minutes of resting data collection, participants cycled at 60 revolutions per minute on an electromagnetically braked cycle ergometer (Lode, Gronigen, The Netherlands) for three minutes in an unloaded "freewheeling" state. A progressively ramped period of exercise (10 to 30 W min^{-1} based on stature, age, and predicted $\dot{V}O_2$)[17] was then undertaken to volitional termination and followed by three minutes recovery. Heart rate (Polar electro, Oy, Finland) was recorded throughout.

Medgraphics Breeze™ software automatically determined $\dot{V}O_2$ peak (defined as the highest $\dot{V}O_2$ during the final 30 seconds of exercise reported), oxygen uptake efficiency slope (OUES), and peak oxygen pulse (O_2 pulse). The AT was manually interpreted by a clinician using the V-slope method,[18] supported by $\dot{V}_E/\dot{V}CO_2$-AT, and $\dot{V}_E/\dot{V}O_2$-AT.

Critical Difference

Natural variation is described by the magnitude of CD and determines the difference in CRF required to demonstrate change not simply due to the "noise" associated with analytical imprecision (represented by CV_A) and biological variation (represented by CV_B), in order for it to be considered clinically meaningful. [13 14] Critical difference uses ANOVA to determine the magnitude of random fluctuation around a homeostatic set point within which there is 95% probability that repeated measures will fall. The 95% probability is represented by a constant k (2.77) in Equation 1 (calculated from $\sqrt{2}$ * 1.96 (two standard deviations)). Coefficients of variation were calculated dividing the standard deviation by the mean score and converted into a percentage as shown in the example of CV_A (Equation 2). The coefficient of analytical variation was subtracted from the CV_W determined from the repeated trials to calculate CV_B (Equation 3).

$$CD = k\sqrt{CV_A^2 + CV_B^2} \qquad \text{(Eq 1)}^{13}$$

Where:

k = constant equal to 2.77 at $P < 0.05$

CV_A = coefficient of analytical variation

CV_B = coefficient of biological variation

CV_A was calculated using the following equation:

$$CV_A = \frac{SD}{\bar{x}} \times 100\ (\%) \qquad \text{(Eq 2)}$$

Where:

SD = standard deviation

$\bar{x}$ = mean

CV_B was calculated from $\dot{V}O_2$ data from each participant, collected at periodic times as described, using the following equation:

$$CV_B = CV_W(\%) - CV_A(\%) \qquad \text{(Eq 3)}$$

Where:

CV_W = coefficient of within participant variation

Consequently, when interpreting CPET results, and in order to address the presence of natural variation, the CD (applied above and below an observed score) must be considered to determine the range in which a patient can present without any change in CRF (ie. before clinical significance can be claimed).

Study arm 2: application of CD metrics to patients

A consecutive sample of 213 patients (Table 1) scheduled for elective colorectal surgery who had undergone CPET testing was retrospectively examined. CPETs were conducted in accordance with the American Thoracic Society/ American College of Chest Physician Statement on Cardiopulmonary Exercise Testing,[19] using identical equipment, investigators, and protocols as outlined in Study arm 1.

Calculated CD metrics were subsequently applied to CPET metrics with established evidence to independently identify unfit patients during pre-surgical assessment.[1-4 6 11 20 21] Reference CRF threshold values were established from the European Association for Cardiovascular Prevention & Rehabilitation (EACPR)/American Heart Association (AHA) Scientific Statement: $\dot{V}O_2$-AT < 11mL O_2 kg^{-1} min^{-1}, $\dot{V}O_2$ peak < 16mL O_2 kg^{-1} min^{-1}, and $\dot{V}_E/\dot{V}CO_2$-AT ≥ 36.[22] The CD for additional CPET metrics was calculated for $\dot{V}_E/\dot{V}O_2$-AT,[3 20 23] and peak O_2 pulse.[5 7 10 24]

To determine the impact of natural variation on fitness stratification, patient counts were calculated for uncorrected (observed) fit and unfit categories according to EACPR/AHA

threshold values, positively corrected (+CD), and negatively corrected (-CD) values. A revised fitness stratification model for each CPET metric was created by applying ±CD to threshold values, thus creating upper and lower boundaries associated with natural variation, and the area in-between the newly defined boundaries classified as indeterminate-fitness. Finally, patient counts were compared for current versus newly revised models.

Statistics

Statistical analyses were conducted using IBM SPSS Statistics for Windows (Version 23.0 Armonk, NY). Distribution normality was confirmed using Shapiro-Wilk *W* tests. Within-subject time of day difference in CPET performance was assessed using Bonferroni corrected repeated measures analysis of variance. Patient counts were analysed using Chi-Square tests. Continuous data are presented as mean (standard deviation) or median (range), and categorical data as absolute values (%). Significance for all two-tailed tests was established at $P < 0.05$. Retrospective sample size calculations were conducted attaining 80% power at the $P < 0.05$ level with the minimum effect of clinical importance represented by the calculated CD (from study arm 1, Table 2) and between-patient standard deviations (from study arm 2, Table 1).[25]

Results

Natural variation

Study arm 1 identified a CD of 19% for $\dot{V}O_2$-AT (CV_A 2.2%, CV_B 6.5%), 13% for $\dot{V}O_2$ peak (CV_A 2.2%, CV_B 3.9%), and 10% for $\dot{V}_E/\dot{V}CO_2$-AT (CV_A 0.6%, CV_B 3.6%) (Table 2.). The time of day that CPET was conducted had no effect in measured metrics ($\dot{V}O_2$-AT: $P = 0.40$, $\dot{V}O_2$ peak: $P = 0.81$, and $\dot{V}_E/\dot{V}CO_2$-AT: $P = 0.75$). When CD was applied to current CPET fitness threshold values of $\dot{V}O_2$-AT: 11mL O_2 kg^{-1} min^{-1}, $\dot{V}O_2$ peak: 16mL kg^{-1} min^{-1}, and $\dot{V}_E/\dot{V}CO_2$-AT: 36, a variation of ±2.1mL O_2 kg^{-1} min^{-1}, ±2.0mL kg^{-1} min^{-1}, and ±3.7 respectively was observed.

Potential for incorrect fitness stratification

We applied CD to positively and negatively correct (the range of) patient CPET scores around their observed (single-point estimate) scores, and subsequently calculated the number of "false positive" and "false negative" results. While these terms are not technically correct given the unavoidable uncertainty associated with biological variation and corresponding inability to determine an individual's "true" level of CRF at any given point in time, it nonetheless provides a conceptual framework to illustrate how blunt application of current thresholds has the potential to affect perioperative planning for a large proportion of patients undergoing major elective surgery.

The application of natural variation (±CD) presented a mathematical possibility for patient results to transcend current fitness stratification boundaries thus demonstrating potential for misclassification (Figure 2) using $\dot{V}O_2$-AT, $\dot{V}O_2$ peak, and $\dot{V}_E/\dot{V}CO_2$-AT ($P < 0.001$ in all cases). Differences in patient counts assigned to a given fitness category resulted in false negatives (whereby patients were stratified as fit with variation positively corrected when they were originally unfit), and false positives (whereby patients were stratified as unfit

with variation negatively corrected when they were originally fit). Thus, natural variation may have caused up to 59 (28%) false negatives and 69 (32%) false positives at the AT, 33 (15%) false negatives and 35 (16%) false positives at peak $\dot{V}O_2$, and 37 (17%) false negatives and 43 (20%) false positives at the $\dot{V}_E/\dot{V}CO_2$-AT.

Revised model

A revised fitness stratification model (Figure 3) was created with CD defining asymmetrical upper and lower boundaries for absolute values (13.6 and 9.2mL O_2 kg^{-1} min^{-1} for AT, 18.3 and 14.2mL kg^{-1} min^{-1} for $\dot{V}O_2$ peak, 40.1 and 32.7 for $\dot{V}_E/\dot{V}CO_2$-AT) that were independent of fitness misclassification based on natural variation. The resultant area between the upper and lower boundaries represented a newly defined and additional category labelled "Indeterminate-fitness". The indeterminate-fitness category accounted for 60, 32, and 40% of patients for the AT, $\dot{V}O_2$ peak and $\dot{V}_E/\dot{V}CO_2$-AT metrics respectively (Figure 4), and thus fewer patients were stratified as unfit or fit.

Discussion

The present findings highlight the potential for incorrect patient fitness stratification when natural variation is not taken into account. We formulated a revised model (accounting for natural variation) which established that many patients were stratified with indeterminate-fitness. We therefore encourage clinicians to be aware of natural variation and its implications for fitness stratification and suggest this concept be applied to markers of CRF to further optimise patient management. Whilst this investigation aims to improve the prognostic interpretation of CPET results, we acknowledge and advocate that clinical decision making does not rely on the application of threshold values alone. There are clear dangers of just using a single point estimate, even if it may be a better number when natural variation is considered. A multitude of additional variables such as work rate, heart rate, duration of exercise, reason for stopping the exercise all go into a composite estimate of functional capacity to be considered alongside other clinical measures when planning perioperative care.

Potential for incorrect patient fitness stratification

The mean CPET score for patients undergoing colorectal surgery was identical to the threshold marker value for AT, within 0.3mL O_2 kg^{-1} min^{-1} for $\dot{V}O_2$ peak, and 2.4 lower for $\dot{V}_E/\dot{V}CO_2$-AT. Thus, when patient scores were positively or negatively corrected with CD, large numbers of patients transcended the EACPR/AHA threshold CRF boundaries demonstrating that natural variation may cause significant rates of incorrect fitness stratification. Of the three primary CPET metrics reported, the AT demonstrated the most incorrectly stratified patients, closely followed by peak $\dot{V}O_2$, and to a lesser albeit significant extent $\dot{V}_E/\dot{V}CO_2$-AT in line with magnitudes of reported CD values and close proximity of patient scores to threshold boundaries. Furthermore, a valid and reliable identification of

$\dot{V}O_2$-AT is not always possible and has been well documented in patients with heart failure,[26] and thus may contribute to greater variance in AT.

Revised fitness stratification

Our revised model (with its wider boundaries accounting for natural variation) excluded many patients from both unfit and fit categories, and thus large numbers were stratified in the indeterminate-fitness category (Figure 4). Not only does this occurrence confirm the impact of natural variation, but consequently presents the challenge of planning perioperative care for patients within this additional fitness category. Concerns may be associated with the introduction of an additional fitness category. For example, patients undergoing colorectal surgery who fell into an intermediate-fitness group (albeit not comparable with our indeterminate-fitness category) have reported a higher rate of serious complications if admitted to the ward rather than HDU.[27]

The most effective way to assess patient risk is likely a combined approach using clinical variables, biomarkers of susceptibility to disease, and physiological testing (CPET).[28] We suggest further development of our model by inclusion of known risk factors independent of CRF to optimise perioperative care.

Limitations

We recognise that this study has limitations and simply reflects a "proof of principle" concept. Measures of CD were derived from young healthy participants and applied to a cohort of older patients. Comparative values for older controls were not available and would present considerable ethical challenges to determine given that repeat CPET to volitional exhaustion would be required. Our CV_W (given by CV_A + CV_B from Table 2) of 6.1% for $\dot{V}O_2$ peak is comparable with chronic obstructive pulmonary disease (6.6%) and congestive

heart failure patients (5.7% and 6.0%).[29-31] Furthermore, our CV_W for AT (8.7%) is consistent with patient data (6.8%, 9.2% and 10%),[32 30 33] and in excess of CV_W values for $\dot{V}O_2$ peak, the probable consequence of observer error when determining AT via the V-slope method.[18] Thus, our method has potential application to clinical populations. However, reported metrics for CD may reflect a best-case scenario (ie. lowest CD) if natural variation increases with age and/or pathology.

Study arm 1 comprised of men only, whilst the calculated CD was subsequently applied to a population of whom 41% were women. For the $\dot{V}O_2$ peak and $\dot{V}O_2$-AT metrics, our coefficients of variation were comparable with the studies previously stated which also included female data. Metrics represented by ventilatory equivalents however must be treated with caution (for female comparison) as any disparity between the sexes is not accounted for.

Many CPET metrics are scaled to body mass. Further investigation is required to determine if there are any effects on the magnitude of asymmetry for absolute values reported around our zones of indeterminate-fitness resulting from scaling to body mass.

Data were collected on a single system in both arms of this study. We are aware that analytical precision is likely to vary widely between different manufacturers thus affecting CV_A and consequently CD. Therefore, our results can only be applied with certainty to clinical tests using Medgraphics equipment. At the time of conducting the study the authors did not have access to a metabolic calibrator used to calculate CV_A however we are confident that our findings (up to 2.2%) are comparable with data produced from such devices which typically report with accuracy of ±2%.[34]

Prospective sample size calculations

From an experimental design perspective, our observations have implications when prospectively determining sample sizes for future randomised controlled exercise trials. We

suggest that CD be used to determine the minimal clinically important difference (MCID) for any given metric of CRF. Until now, studies often rely on MCID values that appear to lack a well-established scientific basis, such as a $\dot{V}O_2$-AT of 2mL kg^{-1} min^{-1} for example.[35] This (arbitrarily) defined MCID of 2mL O_2 kg^{-1} min^{-1} is in fact incorrect because it falls within our calculated CD of 2.1mL O_2 kg^{-1} min^{-1} (i.e. this is part of normal variation). In a worked example using the arbitrary metric of 2mL O_2 kg^{-1} min^{-1}, a prospective power calculation indicates that a two-armed exercise intervention study would require a minimum of 36 patients per group (excluding potential dropout) to detect a treatment effect with 80% power at the $P < 0.05$ level. However, considering natural variation (using our calculated CD of 2.1mL O_2 kg^{-1} min^{-1} in place of 2mL O_2 kg^{-1} min^{-1}) would further inflate the sample size (to 39 patients per group) highlighting the potential for a type II error.

We recognise that the sample size calculation is based upon a CD determined from a sample of 12 subjects and is limited to a single (Medgraphics) system. Further research (with larger sample sizes, additional metabolic carts, and calculations across the spectrum of age, health and CRF) is encouraged to better support our prospective calculation of sample sizes.

Conclusions

These findings demonstrate the extent of natural variation in CPET data. Natural variation also has potential to influence patient fitness stratification. Therefore, clinicians should not consider fitness as a single point estimate, but instead as a dynamic range of values defined by natural variation and calculated using critical difference. We suggest the use of CRF threshold values inclusive of natural variation to optimise risk prediction models, and encourage clinicians to be aware of natural variation and its implications when determining the appropriate level of post-operative care following major surgery.

Author's contributions

All authors were involved in the conception and design of study. R.G.D, I.R.A, G.A.R performed the CPET tests and collated the data. G.A.R. performed the analysis with input from D.M.B, M.H.L, R.G.D, I.R.A. The manuscript was drafted by G.A.R and D.M.B. All authors provided revisions and approved the final version for submission.

Declaration of interest

The authors declare no conflict of interest.

Funding

This work was supported by the Higher Education Funding Council for Wales (to DM Bailey). DM Bailey is a Royal Society Wolfson Research Fellow (#WM170007).

Acknowledgements

Sean Cutler, James O'Flaherty, and Trevor Harris contributed to data collection and technical input in study arm 1.

References

1 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. *Br J Surg* 2016; **103**: 744-52

2 Grant SW, Hickey GL, Wisely NA, et al. Cardiopulmonary exercise testing and survival after elective abdominal aortic aneurysm repair. *Br J Anaesth* 2015; **114**: 430-6

3 Carlisle J, Swart M. Mid-term survival after abdominal aortic aneurysm surgery predicted by cardiopulmonary exercise testing. *Br J Surg* 2007; **94**: 966-9

4 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. *Br J Anaesth* 2013; **111**: 607-11

5 Prentis JM, Trenell MI, Jones DJ, Lees T, Clarke M, Snowden CP. Submaximal exercise testing predicts perioperative hospitalization after aortic aneurysm repair. *J Vasc Surg* 2012; **56**: 1564-70

6 Snowden CP, Prentis J, Jacques B, et al. Cardiorespiratory fitness predicts mortality and hospital length of stay after major elective surgery in older people. *Annals of surgery* 2013; **257**: 999-1004

7 West MA, Parry MG, Lythgoe D, et al. Cardiopulmonary exercise testing for the prediction of morbidity risk after rectal cancer surgery. *Br J Surg* 2014; **101**: 1166-72

8 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**: e653-99

9 Older R, Smith R, Courtney B, Hone R. Preoperative Evaluation of Cardiac Failure and Ischemia in Elderly Patients by Cardiopulmonary Exercise Testing. *Chest* 1993; **104**: 701-4

10 Junejo MA, Mason JM, Sheen AJ, et al. Cardiopulmonary exercise testing for preoperative risk assessment before hepatic resection. *Br J Surg* 2012; **99**: 1097-104

11 Hartley RA, Pichel AC, Grant SW, et al. Preoperative cardiopulmonary exercise testing and risk of early mortality following abdominal aortic aneurysm repair. *Br J Surg* 2012; **99**: 1539-46

12 Moran J, Wilson F, Guinan E, McCormick P, Hussey J, Moriarty J. Role of cardiopulmonary exercise testing as a risk-assessment method in patients undergoing intra-abdominal surgery: a systematic review. *Br J Anaesth* 2016; **116**: 177-91

13 Fraser CG, Fogarty Y. Interpreting laboratory results. *BMJ* 1989; **298**: 1659-60

14 Davison GW, Ashton T, McEneny J, Young IS, Davies B, Bailey DM. Critical difference applied to exercise-induced oxidative stress: the dilemma of distinguishing biological from statistical change. *J Physiol Biochem* 2012; **68**: 377-84

15 Bailey DM, Evans TG, Gower Thomas K. Intervisceral artery origins in patients with abdominal aortic aneurysmal disease; evidence for systemic vascular remodelling. *Exp Physiol* 2016; **101**: 1143-53

16 Williams JR. The Declaration of Helsinki and public health. *Bull World Health Organ* 2008; **86**: 650-2

17 Wasserman K. *Principles of exercise testing and interpretation: including pathophysiology and clinical applications*. 5th ed Edn. London: Wolters Kluwer/Lippincott Williams & Wilkins, 2012

18 Beaver WL, Wassermen K, Whipp BJ. A new method for detecting anaerobic threshold by gas exchange. *J Appl Physiol* 1986; **60**: 2020-7

19 American Thoracic S, American College of Chest P. ATS/ACCP Statement on cardiopulmonary exercise testing. *Am J Respir Crit Care Med* 2003; **167**: 211-77

20 West MA, Loughney L, Barben CP, et al. The effects of neoadjuvant chemoradiotherapy on physical fitness and morbidity in rectal cancer surgery patients. *Eur J Surg Oncol* 2014; **40**: 1421-8

21 Wilson RJ, Davies S, Yates D, Redman J, Stone M. Impaired functional capacity is associated with all-cause mortality after major elective intra-abdominal surgery. *Br J Anaesth* 2010; **105**: 297-303

22 Guazzi M, Arena R, Halle M, Piepoli MF, Myers J, Lavie CJ. 2016 Focused Update: Clinical Recommendations for Cardiopulmonary Exercise Testing Data Assessment in Specific Patient Populations. *Circulation* 2016; **133**: e694-711

23 Junejo MA, Mason JM, Sheen AJ, et al. Cardiopulmonary exercise testing for preoperative risk assessment before pancreaticoduodenectomy for cancer. *Ann surg oncol* 2014; **21**: 1929-36

24 Epstein SK, Freeman RB, Khayat A, Unterborn JN, Pratt DS, Kaplan MM. Aerobic capacity is associated with 100-day outcome after hepatic transplantation. *Liver transpl* 2004; **10**: 418-24

25 Altman DG. Statistics and ethics in medical research: III How large a sample? *BMJ* 1980; **281**: 1336-8

26 Arena R, Myers J, Williams MA, et al. Assessment of functional capacity in clinical and research settings: a scientific statement from the American Heart Association Committee on Exercise, Rehabilitation, and Prevention of the Council on Clinical Cardiology and the Council on Cardiovascular Nursing. *Circulation* 2007; **116**: 329-43

27 Swart M, Carlisle JB, Goddard J. Using predicted 30 day mortality to plan postoperative colorectal surgery care: a cohort study. *Br J Anaesth* 2017; **118**: 100-4

28 Grocott MPW. Improving outcomes after surgery. *BMJ* 2009; **339**: b5173

29 Owens MW, Kinasewitz GT, Strain DS. Evaluating the Effects of Chronic Therapy in Patients with Irreversible Air-Flow Obstruction. *Am Rev Respir Dis* 1986; **134**: 935-7

30 Janicki JS, Gupta S, Ferris ST, McElroy PA. Long-term Reproducibility of Respiratory Gas Exchange Measurements during Exercise in Patients with Stable Cardiac Failure. *Chest* 1990; **97**: 12-7

31 Elborn JS, Stanford CF, Nicholls DP. Reproducibility of cardiopulmonary parameters during exercise in patients with chronic cardiac failure. The need for a preliminary test. *Eur Heart J* 1990; **11**: 75-81

32 Keteyian SJ, Brawner CA, Ehrman JK, Ivanhoe R, Boehmer JP, Abraham WT. Reproducibility of peak oxygen uptake and other cardiopulmonary exercise parameters: implications for clinical trials and clinical practice. *Chest* 2010; **138**: 950-5

33 Kothmann E, Danjoux G, Owen SJ, Parry A, Turley AJ, Batterham AM. Reliability of the anaerobic threshold in cardiopulmonary exercise testing of patients with abdominal aortic aneurysms. *Anaesthesia* 2009; **64**: 9-13

34 Huszczuk A, Whipp B, Wasserman K. A respiratory gas exchange simulator for routine calibration in metabolic studies. *Eur Respir J* 1990; **3**: 465-8

35 Kothmann E, Batterham AM, Owen SJ, et al. Effect of short-term exercise training on aerobic fitness in patients with abdominal aortic aneurysms: a pilot study. *Br J Anaesth* 2009; **103**: 505-10

Table 1. **Participant and patient characteristics.** Data are shown as mean (± standard deviation) or ~(range), and *n (%). n, number; IHD, ischaemic heart disease; COPD, chronic obstructive pulmonary disease; $\dot{V}O_2$ peak, peak oxygen consumption; RER, respiratory exchange ratio; AT, estimated anaerobic threshold; $\dot{V}_E/\dot{V}CO_2$, ventilatory equivalent for carbon dioxide; $\dot{V}_E/\dot{V}O_2$, ventilatory equivalent for oxygen; O_2 pulse, oxygen pulse at peak exercise; Work load at AT, work load at estimated anaerobic threshold; Workload at peak, work load at peak exercise.

	Study arm 1 Apparently healthy participants (n = 12)	***Study arm 2*** Colorectal patients (n = 213)
Demographics:		
Age (years) ~	22 (20-26)	69 (32-90)
BMI	26 (3.1)	28.3 (5.8)
Sex*		
male	12 (100)	126 (59)
female	0 (0)	87 (41)
Risk factors:		
Smoking*		
no	12 (100)	71 (33)
yes (active/former)	0 (0)	142 (67)
Hypertension*	0 (0)	79 (37)
Diabetes*	0 (0)	34 (16)
IHD*	0 (0)	37 (17)
COPD*	0 (0)	21 (10)
Haemoglobin (g L^{-1})	-	12.7 (1.9)
Creatinine (µmol L^{-1})	-	79.2 (19.7)
Cardiopulmonary function:		
Baseline heart rate (beats min^{-1})	65 (5)	83 (19)
Peak heart rate (beats min^{-1})	178 (5)	124 (28)
$\dot{V}O_2$ peak (mL kg^{-1} min^{-1})	43.8 (6.0)	16.3 (4.9)
RER at peak $\dot{V}O_2$	1.3 (0.1)	1.1 (0.1)
AT (mL O_2 kg^{-1} min^{-1})	23.8 (3.6)	11.0 (3.0)
$\dot{V}_E/\dot{V}CO_2$-AT	23.5 (1.4)	33.6 (5.3)
$\dot{V}_E/\dot{V}O_2$-AT	23.5 (4.7)	30.6 (5.9)
O_2 pulse (mL $beat^{-1}$)	20.7 (0.9)	10.5 (3.8)
Work load at AT (W)	160 (28)	52 (28)
Work load at peak (W)	300 (45)	91 (47)

Table 2. Biological variation and critical difference for cardiopulmonary exercise test variables (Study arm 1, n=12). CV_A, coefficient of analytical variation; CV_B, coefficient of biological variation; AT, anaerobic threshold; $\dot{V}O_2$ peak, peak oxygen consumption; $\dot{V}_E/\dot{V}CO_2$, ventilatory equivalent for carbon dioxide; $\dot{V}_E/\dot{V}O_2$, ventilatory equivalent for oxygen; O_2 pulse, oxygen pulse at peak exercise; OUES, oxygen uptake efficiency slope; RER, respiratory exchange ratio.

Parameter	CV_A (%)	CV_B (%)	Critical difference (%)
AT (mL O_2 kg^{-1} min^{-1})	2.2	6.5	19.1
$\dot{V}O_2$ peak (mL kg^{-1} min^{-1})	2.2	3.9	12.5
$\dot{V}_E/\dot{V}CO_2$-AT	0.6	3.6	10.2
$\dot{V}_E/\dot{V}O_2$-AT	1.7	3.0	9.6
O_2 pulse (mL $beat^{-1}$)	2.2	2.3	8.9
OUES	2.2	3.8	12.1
RER at peak exercise	1.4	5.3	15.2

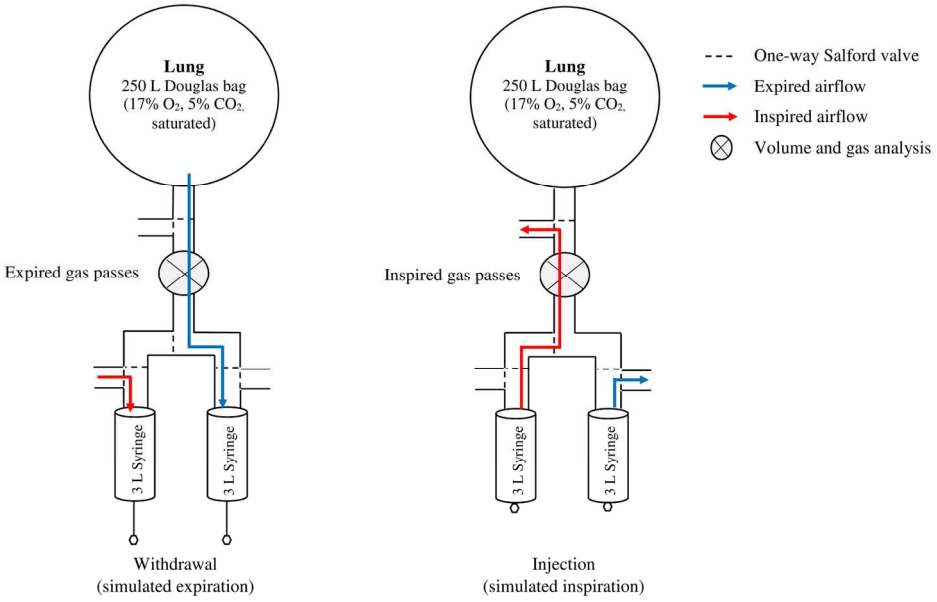

Figure 1. The determination of CV_A for CPET metrics using simulated expiration and inspiration. CV_A, analytical coefficient of variation; CPET, cardiopulmonary exercise test. Simulated oxygen uptake for trials ~ 13mL kg^{-1} min^{-1}.

153x98mm (300 x 300 DPI)

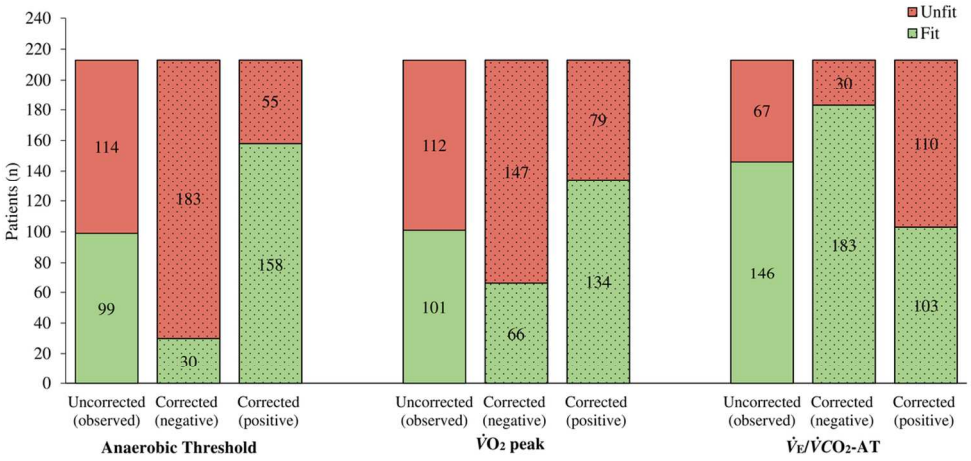

Figure 2. Potential for incorrect patient fitness stratification if natural variation is not taken into account. Patient counts are presented for unfit (AT < 11mL O_2 kg^{-1} min^{-1}, VO_2 peak < 16mL kg^{-1} min^{-1}, $V_E/VCO_2 \geq 36$) and fit (AT ≥ 11mL O_2 kg^{-1} min^{-1}, peak $VO_2 \geq$ 16mL kg^{-1} min^{-1}, $V_E/VCO_2 < 36$) categories. AT, anaerobic threshold; VO_2 peak, peak oxygen consumption; V_E/VCO_2, ventilatory equivalent for carbon dioxide; Observed, uncorrected scores indicative of current risk stratification; Positive, corrected scores by addition of CD; Negative, corrected scores by subtraction of CD. P < 0.001 across all pairwise comparisons for corrected scores. Natural variation caused 59 (28%) false negatives and 69 (32%) false positives at the AT, 33 (15%) false negatives and 35 (16%) false positives at VO_2 peak, and 37 (17%) false negatives and 43 (20%) false positives at the V_E/VCO_2-AT.

114x53mm (300 x 300 DPI)

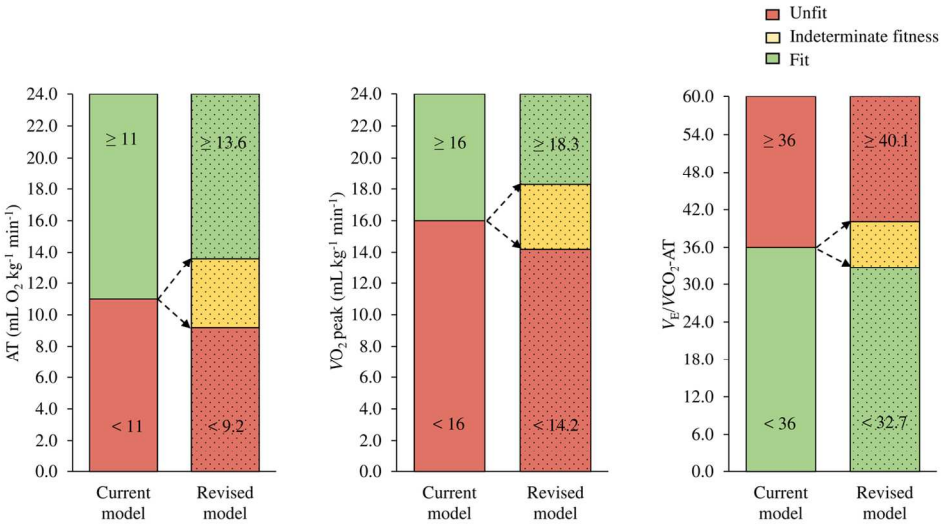

Figure 3. Revised fitness stratification model following incorporation of the critical difference for the anaerobic threshold, $\dot{V}O_2$ peak, and $\dot{V}_E/\dot{V}CO_2$-AT. AT, anaerobic threshold; $\dot{V}O_2$ peak, peak oxygen consumption; $\dot{V}_E/\dot{V}CO_2$, ventilatory equivalent for carbon dioxide. Natural variation demonstrates the magnitude of variation present. The lower and upper boundaries define clinically meaningful boundaries not affected by natural variation whilst the area in-between is classified as indeterminate fitness.

131x72mm (300 x 300 DPI)

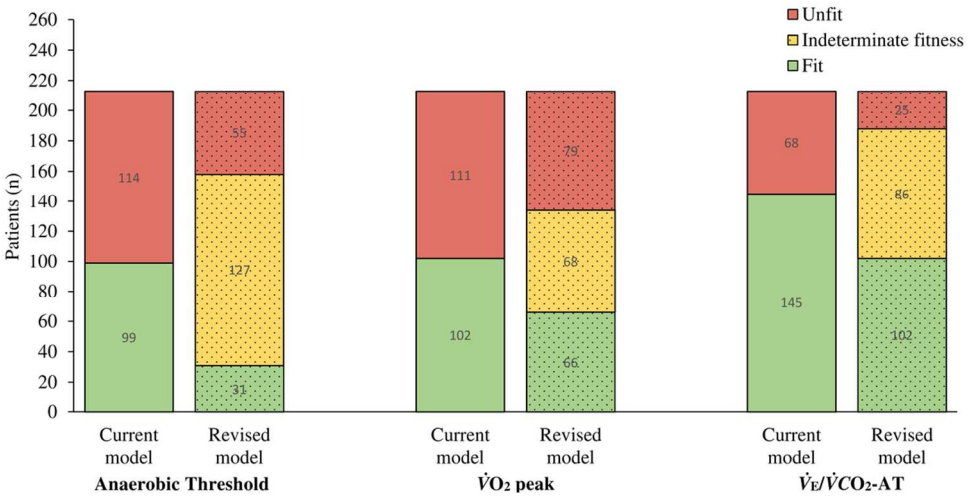

Figure 4. Current versus revised model identification of patient counts by fitness category. AT, anaerobic threshold; $\dot{V}O_2$ peak, peak oxygen consumption; $\dot{V}_E/\dot{V}CO_2$, ventilatory equivalent for carbon dioxide. The revised model demonstrates large numbers of patients that are classified with indeterminate fitness.

111x56mm (300 x 300 DPI)