Cardiopulmonary exercise testing predicts 5 yr survival after major surgery†

M. Colson1*, J. Baglin2, S. Bolsin1 and M. P. W. Grocott3

1 Department of Anaesthesia, The Geelong Hospital, Ryrie Street, Geelong, VIC 3220, Australia
2 School of Mathematics and Geospatial Sciences, RMIT University, Plenty Road, Bundoora, VIC 3083, Australia
3 Southampton University Hospital NHS Trust, Tremona Road, Southampton SO16 6YD, UK
* Corresponding author. E-mail: markcolsonline@gmail.com

Editor’s key points
- Database of 1725 patients with cardiopulmonary exercise testing (CPET) who had major surgery.
- Five-year survival was not associated with any single variable.
- Multivariate modelling identified gender, type of surgery, and forced vital capacity as major predictors.
- Addition of CPET variables derived at anaerobic threshold improved the accuracy.

Background. Cardiopulmonary exercise testing (CPET) is used to assess perioperative risk in surgical patients. While previous studies have looked at short-term outcomes, this paper explores the ability of CPET to predict 5 yr survival after major surgery.

Methods. Over a period (1996–2009), 1725 patients referred for CPET subsequently underwent major surgery. Breath-by-breath data derived during each patient’s CPET was processed using customized software to extract variables likely to impact on survival. Initial analysis examined the predictive power of single variables. Subsequently, Bayesian model averaging (BMA) was used to construct a multivariate model defining the association between CPET data and 5 yr survival.

Results. Six hundred and sixteen (36%) of the study patients died. Single variables were not significantly associated with 5 yr postoperative survival. BMA indicated the following major predictors of 5 yr survival: patient gender; type of surgery, and forced vital capacity. Four variables derived at the patient’s anaerobic threshold were weaker predictors. These were end-tidal oxygen concentration, respiratory exchange ratio, oxygen consumption per unit body weight, and oxygen consumption per heart beat. The resulting model was then used to divide patients into low-, medium-, or high-risk categories, and 5 yr survival for each category was 87.8; 75.8, and 53.8% respectively. Survival was independent of patient age.

Conclusions. Multivariate analysis and model generation techniques can be applied to CPET data to predict 5 yr survival after major surgery more accurately than is possible with single variable analysis.

Keywords: anaesthesia, audit; lung, gas exchange, respiratory; statistics; surgery, non-cardiac; surgery, postoperative period

Accepted for publication: 21 May 2012

Preoperative assessment of patients for non-cardiac surgery has taxed anaesthetists for many years.1 Scoring systems have been developed along with guidelines for risk assessment, specifically cardiac risk.2–4 Cardiopulmonary exercise testing (CPET) has been successfully used to improve the accuracy of preoperative prediction of perioperative complications and survival.5–7 This work emphasized the anaerobic threshold (AT) as the definitive CPET endpoint. While the predictive power of the oxygen consumption at AT (AT-V\textsubscript{O2}) has been confirmed by subsequent research,8 9 the primacy of this single variable in the context of perioperative risk prediction has been questioned.10 Other groups have broadened the search for predictive information derived from a CPET test to include such variables as the ventilatory equivalents11 and oxygen uptake efficiency slope (OUES) (Table 1).12

Our institution began CPET of patients before major surgery in 1996. The primary aim of this study was to explore the relationship between CPET-derived variables and 5 yr survival after elective major surgery. We hypothesized that statistical modelling using multiple CPET variables would discriminate between survivors and non-survivors better than any single CPET variable.

Methods
The data collection required for this project received institutional approval.13 Written informed consent for data collection was obtained from each patient in the study. No patient declined to consent to their data being collected.

†A preliminary version of this work was presented in poster format at the Australia and New Zealand College of Anaesthetists’ Annual Scientific Meeting in Hong Kong in May 2011.

© The Author [2012]. Published by Oxford University Press on behalf of the British Journal of Anaesthesia. All rights reserved. For Permissions, please email: journals.permissions@oup.com
Patients over 60 yr of age undergoing elective major abdominal or thoracic surgery who had no contraindications were offered CPET as part of our institutional protocol. In addition, younger patients were often referred for CPET testing based on findings at preoperative consultation. Contraindications to CPET included unstable coronary syndromes, orthopaedic or vascular conditions which precluded pedalling, and an inability to breathe through a mouthpiece.

CPET was conducted in the days (median 9 days) before surgery. Patients fasted for 4 h before the test, which was supervised by a physician and nurse technician. Twelve-lead ECG analysis was performed throughout each test. MedGraphics (Medical Graphics Corporation, St Paul, MN, USA) exercise testing equipment was used which was updated several times through the study period. The most recent installation was a Lode electromagnetically braked bike (Lode BV, Groningen, The Netherlands) connected to BreezeSuite 7.0 software (Medical Graphics Corporation) running on an MS Windows 7 platform (Microsoft Corporation, Seattle, Washington, DC, USA). Each test used a graded exercise protocol in which each patient’s AT workload was estimated in advance, and the rate of load increment selected so as to reach this workload within 10 min. Each test began with a 3 min unloaded section, after which the load continuously increased at the predetermined rate. The test was terminated when the supervising physician felt confident that the AT had been reached. In some cases, tests were terminated for other reasons, including patient intolerance, or the development of ECG changes consistent with myocardial ischaemia. Patients who did not reach their AT are not included in the analysis.

Breath-by-breath test data were exported into customized software developed to extract 42 variables of interest from each CPET. These include biometric and patient variables, plus a number of physiological variables at two stages of the test. The chosen stages were the unloaded section of the test (where each variable was averaged over the full unloaded section) and the AT. The maximal load point was deliberately not used as a test juncture, since this did not occur at a consistent, reproducible point in each test. AT was determined using an automated V-slope algorithm developed by the author and based on the original description of the technique.14 The results of this technique showed near-perfect concordance with the automated V-slope detection method included in the BreezeSuite software. These data were then subject to statistical analysis.

With regard to clinical outcomes, 6 months after each CPET, the patient’s case notes were reviewed and the details of any major surgical procedures performed in our institution during this period were logged. Patient death data were obtained from the institution’s electronic patient information system, which in turn was cross-referenced with a national death database. This technique has been validated as being highly accurate. Although death data were also available for those patients who had not undergone any major surgery within our institution during the study period, it was decided to eliminate this cohort from the analysis, as many of these patients may have in fact undergone surgery in other jurisdictions where we did not have access to their perioperative information. Death data were censored on a date after the end of the study period. By using the Cox proportional hazard (PH) statistical technique, which handles censored survival data, data derived from patients whose surgery was <5 yr from the censoring date were nevertheless able to be included in the analysis.

To control for patient prognosis following different types of surgery, patients were categorized into seven major surgery types. These were ranked in order of their survival from 1 (highest mean survival days) to 7 (lowest mean survival days). This rank was used as a variable in the statistical modelling. Many of the patients underwent cancer surgery, but this was not included in the model as in many cases, cancer status was not accurately ascertained until during or after surgery, and was, therefore, of limited value in preoperative risk prediction.

### Statistical analysis

#### Single variable analysis

A number of single variables derived during CPET have been reported to be associated with outcomes after major surgery, including the AT and the ventilatory equivalents for carbon dioxide at AT (AT-V̇E/V̇CO₂).9 The OUES has been proposed as
an independent measure of cardiorespiratory function which is especially useful in submaximal CPET. These three variables, and patient age, were examined individually to determine their association with three survival endpoints: 30-day mortality, 1 yr mortality, and 5 yr mortality. Receiver-operator characteristic (ROC) curves were constructed for each of the resulting 12 CPET-variable–endpoint combinations.

Multivariate analysis

PH models were used to model 5 yr postoperative survival using patient and CPET variables for each patient. PH models are an extension of logistic regression which takes into account the time to an event (in this case, patient death). Given the large number of potential inputs into the predictive model, the decision was made a priori that Bayesian model averaging (BMA) would be used to avoid the problem of arbitrary input selection and resulting model instability. The BMA approach accounts for the effects of model uncertainty. Model uncertainty arises from not knowing the true population statistical model. In variable selection model building problems, standard approaches, such as forward and backwards stepwise selection, do not account for model uncertainty. Standard approaches choose the best model from the data and then proceed as if that model has produced the data. This may result in underestimated model parameters, overestimated confidence in the selected model being true, riskier decision-making, or poor predictive performance.

Given the large number of CPET predictor variables available for model building, a high degree of model uncertainty was anticipated. To deal with this uncertainty, a BMA method for PH models developed by Volinsky and colleagues was utilized and implemented using a BMA application developed for the statistical package R. BMA addresses model uncertainty by averaging model parameter estimates across a reduced set of best fitting models known as Occam’s Window. Occam’s Window was defined by models falling within 5% of the posterior probability of the best model. The best model was selected based on the largest Bayesian information criteria of the model with the highest posterior probability. A leaps-and-bounds algorithm was used to efficiently target models within 5% of the posterior probability of the best model.

The models occupying Occam’s Window were then fitted to the data to obtain estimates of the PH coefficients. These coefficients were then averaged over all models in Occam’s Window using the posterior model probabilities as weights. From these averaged estimates, the posterior probability that a coefficient is non-zero was calculated. The posterior probability, \( P(B \neq 0) \), is interpreted as the probability that a predictor has an effect. The interpretation of \( P(B \neq 0) \) has been categorized: \(<50\%\) is evidence against an effect; \(50–75\%\) is weak evidence; \(75–95\%\) is positive evidence; \(95–99\%\) is strong evidence; and \(>99\%\) is very strong evidence of an effect.

The performance of the BMA approach was assessed by measuring the proportion of non-survivors allocated to each risk category as defined by a tertile split of the model’s calculated risk score.

Results

Data collection began on 1 January 1996 and continued through to 31 December 2009. During this 14 yr period, 2256 patients underwent 2317 CPET tests. A total of 1725 patients proceeded to major surgery during the trial period and were included in this study. No analysis of the non-surgical cohort was performed. The patient population was predominantly Caucasian with 65% male. Patient ages ranged from 36 to 93 (median 71) yr.

Single variable analysis

The area under the ROC curve (AUC) value was \(<0.65\) for all 12 variable–endpoint combinations (Table 2). Thus, the association between any of these variables and endpoints was not sufficiently strong to warrant further analysis.

Multivariate analysis

Patients were categorized into seven surgical types, and the survival for each type plotted (Fig. 1). Upper gastrointestinal surgery was associated with the worst survival and vascular surgery the best survival. There was a decrease in survival for

<table>
<thead>
<tr>
<th>Table 2</th>
<th>Single variable analysis. Area under the ROC curve linking the sensitivity and specificity of selected single CPET variables in predicting survival endpoints</th>
</tr>
</thead>
<tbody>
<tr>
<td>Outcome</td>
<td>AT-V_\text{O}_2, \text{kg}^{-1}</td>
</tr>
<tr>
<td>30-day mortality</td>
<td>0.551</td>
</tr>
<tr>
<td>One-year mortality</td>
<td>0.474</td>
</tr>
<tr>
<td>Five-year mortality</td>
<td>0.502</td>
</tr>
</tbody>
</table>

Fig 1 Survival by surgery type. The Kaplan–Meier 5 yr survival curves across the seven major surgery types.
the vascular group after 4.5 yr, resulting in the vascular group having higher mean survival days, but lower 5 yr survival than the urological group.

The Kaplan–Meier survival estimates ordered by mean days survived (M) for the number of patients (n) in each surgery type were as follows: vascular, n = 129, M = 1609 [95% confidence interval (CI), 1520–1698], 79% survival; urological, n = 302, M = 1576 (95% CI, 1513–1639), 82% survival; miscellaneous, n = 275, M = 1496 (95% CI, 1425–1566), 74% survival; major abdominal, n = 568, M = 1422 (95% CI, 1368–1475), 70% survival; intermediate abdominal, n = 191, M = 1370 (95% CI, 1273–1467), 67% survival; thoracic, n = 154, M = 1280 (95% CI, 1171–1390), 60% survival; upper gastrointestinal, n = 97, M = 1120 (95% CI, 971–1270), 52% survival.

The BMA analysis accounted for model uncertainty by averaging parameters over the 354 models identified in Occam’s Window. Three predictors were associated with a 100% posterior probability of an associated effect: gender, surgery type, and forced vital capacity ratio (FVCR) (Table 3). The efficacy of the BMA model was illustrated by the fact that 46.2% of those categorized as being high risk were non-survivors (Table 4). Patient age as a predictor of long-term survival was not supported by BMA. The posterior probability for evidence of patient age influencing survival was weak [P(B $\neq$ 0) = 34%].

Although the three most powerful predictors were not derived from the CPET itself, a number of CPET variables did demonstrate strong evidence of an effect after BMA analysis, including AT-$P_{E_{O2}}$ [P(B $\neq$ 0) = 70%], AT-$V_{O2}$/HR [P(B $\neq$ 0) = 65%], AT-RER [P(B $\neq$ 0) = 57%], and AT-$V_{O2}$,kg$^{-1}$ [P(B $\neq$ 0) = 54%].

**Discussion**

We found weak associations between commonly used CPET variables and 5 yr survival. However, applying sophisticated statistical modelling techniques to the CPET test data, we were able to discriminate between low, medium, and high risk of mortality in the 5 yr period after major surgery.

The BMA analysis demonstrates the effects of gender, surgery type, and FVCR on 5 yr survival after major surgery. Clearly, the determination of these three variables does not require CPET. The predictive power of gender to predict survival is no doubt reflected in the well-recognized longevity of females over males. The FVCR confirms the importance of static lung volume—as a ratio of the predicted value. The forced expiratory volume in 1 s (FEV1) did not appear as a predictor. This could be explained by all the predictive value of the FEV1 being contained in one of the other predictors which did rise to high significance. The next most valuable variables included CPET data and were AT-$P_{E_{O2}}$, AT-$V_{O2}$/HR, AT-RER, and AT-$V_{O2}$,kg$^{-1}$. Patient age did not appear to be an important predictor once survival was adjusted for gender, surgery type, and selected CPET variables. This implies that CPET data do not need to be age-adjusted, as has been suggested by some authors. A good and RER at the AT—perhaps because of anxiety-related mild hyperventilation before reaching their AT—had worse survival. As the BMA process analyses all variables simultaneously, it is inappropriate to analyse any one of these variables in isolation, with respect to its relationship with outcome. This is because of a degree of covariance between closely related variables such as AT-$V_{O2}$/HR and AT-$V_{O2}$,kg$^{-1}$. The strength of the BMA approach is that each predictor’s contribution to the model is weighed according to its importance—this weight being reflected in the coefficient for that variable. As BMA accounts for model uncertainty, the conclusions drawn from a BMA analysis should be the most robust. A robust model is essential if it is to be used to make future predictions. Thus, the BMA model was chosen to construct the Kaplan–Meier survival curves.

---

**Table 3** Statistical analysis. PH model coefficient estimates derived using BMA. Only predictors where P(B $\neq$ 0) > 0.05 are included (gender is defined as male = 1, female = 0; surgery type is as defined in Fig. 1).

<table>
<thead>
<tr>
<th>Predictor</th>
<th>$M_s$</th>
<th>P(B $\neq$ 0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>0.7400</td>
<td>1.00</td>
</tr>
<tr>
<td>FVCR</td>
<td>−1.3850</td>
<td>1.00</td>
</tr>
<tr>
<td>Surgery type</td>
<td>0.2250</td>
<td>1.00</td>
</tr>
<tr>
<td>AT-$P_{E_{O2}}$</td>
<td>0.1140</td>
<td>0.70</td>
</tr>
<tr>
<td>AT-$V_{O2}$/HR</td>
<td>−0.0860</td>
<td>0.65</td>
</tr>
<tr>
<td>AT-RER</td>
<td>−1.2000</td>
<td>0.57</td>
</tr>
<tr>
<td>AT-$V_{O2}$,kg$^{-1}$</td>
<td>0.0440</td>
<td>0.54</td>
</tr>
<tr>
<td>Age</td>
<td>0.0050</td>
<td>0.34</td>
</tr>
<tr>
<td>AT-$P_{E_{O2}}$</td>
<td>−0.0690</td>
<td>0.30</td>
</tr>
<tr>
<td>US-$V_{CO2}$</td>
<td>0.0110</td>
<td>0.29</td>
</tr>
<tr>
<td>US-$V_{O2}$/HR</td>
<td>−0.0610</td>
<td>0.27</td>
</tr>
<tr>
<td>US-$V_{O2}$,kg$^{-1}$</td>
<td>0.0330</td>
<td>0.26</td>
</tr>
<tr>
<td>AT-$V_{E}$/V$CO2$</td>
<td>0.0060</td>
<td>0.21</td>
</tr>
<tr>
<td>US-$P_{E_{O2}}$</td>
<td>0.0360</td>
<td>0.18</td>
</tr>
<tr>
<td>BMI</td>
<td>−0.0050</td>
<td>0.17</td>
</tr>
<tr>
<td>Overall-IR</td>
<td>0.3760</td>
<td>0.17</td>
</tr>
<tr>
<td>US-RER</td>
<td>−0.3900</td>
<td>0.16</td>
</tr>
<tr>
<td>AT-$V_{E}$</td>
<td>0.0030</td>
<td>0.11</td>
</tr>
<tr>
<td>OUES</td>
<td>−0.0020</td>
<td>0.10</td>
</tr>
<tr>
<td>US-HR</td>
<td>0.0010</td>
<td>0.09</td>
</tr>
<tr>
<td>AT-$V_{E}$/V$O2$</td>
<td>−0.0020</td>
<td>0.08</td>
</tr>
<tr>
<td>US-IR</td>
<td>1.1020</td>
<td>0.08</td>
</tr>
<tr>
<td>AT-TV</td>
<td>−0.0130</td>
<td>0.05</td>
</tr>
</tbody>
</table>

**Table 4** Model performance. Risk assignment performance (in predicting 5 yr survival) for each risk category using a PH model and BMA.

<table>
<thead>
<tr>
<th>Censored (%)</th>
<th>Deceased (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>447 (87.8)</td>
</tr>
<tr>
<td>Medium</td>
<td>385 (75.8)</td>
</tr>
<tr>
<td>High</td>
<td>273 (53.8)</td>
</tr>
<tr>
<td>Total</td>
<td>1105 (72.5)</td>
</tr>
</tbody>
</table>

Downloaded from https://academic.oup.com/bja/article-abstract/109/5/735/306041 by guest on 07 November 2018
incidence of the primary endpoint (mortality). We analysed all the data derived during CPET testing with the aim of measuring the contribution of each variable to the primary endpoint. We believe that the tendency to ascribe definitive endpoint status to single CPET variables (such as AT) is not justified. This approach is predominantly based on that used by early investigators in the field and fails to utilize the rich variety of data obtained during CPET. A typical test can easily generate 50,000 data points. It is most unlikely that all or even most of the meaningful predictive value of a CPET test could be summarized in one value of a single variable.

Potential weaknesses of this study include the single-centre retrospective design, the inevitability of a degree of selection bias, automated analysis of CPET test data, the un-blinded design, the use of submaximal exercise testing, and the absence of known patient co-morbidities in the predictive model.

The data are derived from a single-centre and therefore may not apply to other centres. Furthermore, the data were derived over a long time period, over which equipment, techniques, personnel, and other factors may have changed. The analysis is retrospective which increases the possibility of bias. As these CPET tests were conducted as part of a routine clinical service, clinicians were not blinded to the results. The relationship between predictor variables and outcomes may therefore have been obscured by confounding by indication. The use of automated software to analyse the tests could be criticized as others have emphasized the importance of subjective interpretation to achieve high reliability for CPET variables, particularly the AT. However, we consider that this strengthens the results, as we have removed any subjective element from the analysis process.

The mathematical processing of the data is highly complex and involved a long sequence of manual and automated steps, and the development of customized computer software. Inevitably, such complexity can introduce subtle errors into the results.

The use of a submaximal testing protocol could be a potential limitation. Throughout the study period, our centre conducted submaximal exercise tests for perioperative patients. There is no published evidence showing that maximal exercise testing generates more predictive data than does submaximal exercise testing in the perioperative setting. Accordingly, we have used submaximal testing for patient comfort, and possibly safety. The OUES is recognized as correlating with $V_{O2}$ max, so although we did not measure $V_{O2}$ max, the OUES might be considered as a surrogate. As OUES did not appear to be a strong predictor, this would suggest that there is little further predictive information to be gained from the maximal exercise test. However, our project was not designed to address this question.

A deliberate decision was made to avoid the inclusion of any known patient co-morbidities into the predictive model. We believe that to do so may have introduced an additional layer of subjective information. This was inconsistent with our aim of a risk prediction based solely on information.
derived during a uniform exercise test which is nearly immune to subjective factors such as patient recall, education level, and even language barriers.

We found that AT had more limited predictive ability than that shown in previous perioperative studies. This may be due, at least in part, to the relationship between short- and long-term outcome variables. For example, our data show the ROC AUC for AT-V\textsubscript{O}_2 kg\textsuperscript{-1} and 30-day mortality to be 0.551, whereas for 5 yr mortality, it is 0.502.

The validity of our findings needs to be tested prospectively in a different patient population. The authors hope to be able to contribute technical assistance, including automated analysis software, to other groups seeking to replicate this study using their own data, and welcome such approaches.

In conclusion, this study adds three main findings to our knowledge of the subject of CPET testing. First, single-variable primary endpoints commonly derived from CPET tests (such as AT-V\textsubscript{O}_2 kg\textsuperscript{-1}) did not predict 5 yr survival in our analysis when used alone. Secondly, multivariate analysis using a large number of CPET and non-CPET variables did predict outcome in the longer term. Although more complex, such an approach lends itself to automation. Finally, the information derived from CPET testing is superior to patient age as a survival predictor. CPET testing yields a large amount of data which contain previously unrecognized and valuable predictive information. This information can be used to ascribe long-term survival risk to major surgical patients. We believe that it constitutes an important addition to the scientific evidence base linking exercise testing to perioperative medicine.

Acknowledgements

M.C. undertook sabbatical leave during 2010 during which he was based at the Southampton University Hospital NHS Trust, Hampshire, UK, where he worked with M.P.W.G. Much of the analysis work for this study occurred during this placement.

Declaration of interest

M.P.W.G. has received honoraria and travel expenses for two lectures at international congresses from Cortex Biopysik GmbH who manufacture CPET equipment.

References

13. Barwon Health Research and Ethics Advisory Committee, Project No. 97/26


29 Grocott MPW, Pearse RM. Prognostic studies of perioperative risk: robust methodology is needed. *Br J Anaesth* 2010; **105**: 243–5