INTERVENTIONAL CARDIOLOGY AND SURGERY

Validation of four different risk stratification systems in patients undergoing off-pump coronary artery bypass surgery: a UK multicentre analysis of 2223 patients

S Al-Ruzzeh, G Asimakopoulos, G Ambler, R Omar, R Hasan, B Fabri, A El-Gamel, A DeSouza, V Zamvar, S Griffin, D Keenan, U Trivedi, M Pullan, A Cale, M Cowen, K Taylor, M Amrani

Heart 2003;**89**:432–435

See end of article for authors' affiliations

Correspondence to: Mohamed Amrani, Harefield Hospital, Middlesex UB9 6JH, UK; mr.amrani@rbh.nthames.nhs.uk

Accepted 20 November 2002

Background: Various risk stratification systems have been developed in coronary artery bypass graft surgery (CABG), based mainly on patients undergoing procedures with cardiopulmonary bypass.
Objective: To assess the validity and applicability of the Parsonnet score, the EuroSCORE, the American College of Cardiology/American Heart Association (ACC/AHA) system, and the UK CABG Bayes model in patients undergoing off-pump coronary artery bypass surgery (OPCAB) in the UK.
Methods: Data on 2223 patients who underwent OPCAB in eight cardiac surgical centres were collected. Predicted mortality risk scores were calculated using the four systems and compared with observed mortality. Calibration was assessed by the Hosmer–Lemeshow (HL) test. Discrimination was assessed using the receiver operating characteristic (ROC) curve area.
Results: 30 of 2223 patients (1.3%) died in hospital. For the Parsonnet score the HL test was significant ($p < 0.001$) and the receiver operating characteristic curve (ROC) area was 0.74. For the EuroSCORE the HL test was also significant ($p = 0.008$) and the ROC area was 0.75. For the ACC/AHA system the HL test was non-significant ($p = 0.7$) and the ROC area was 0.75. For the UK CABG Bayes model the HL test was also non-significant ($p = 0.3$) and the ROC area was 0.81.
Conclusions: The UK CABG Bayes model is reasonably well calibrated and provides good discrimination when applied to OPCAB patients in the UK. Among the other three systems, the ACC/AHA system is well calibrated but its discrimination power was less than for the UK CABG Bayes model. These data suggest that the UK CABG Bayes model could be an appropriate risk stratification system to use for patients undergoing OPCAB in the UK.

The value of risk stratification derives from the worldwide interest in healthcare outcomes and, more importantly, outcome driven quality assessment.[1] By objectively stratifying patients according to the severity of their disease, risk models provide important tools to analyse health outcomes retrospectively, to identify and quantify the effects of changes of techniques or management, and to enable valid comparisons to be made across time between nations, institutions, and even individual surgeons.[2] Furthermore, risk models can detect changes or differences in risk profiles,[3] help plan effective and optimal use of limited healthcare resources,[4] and most importantly make the process of "informed consent" more feasible and more ethical.[5]

The Parsonnet system was developed in the USA and is probably the most widely used method of "stratifying open heart operations into levels of predicted operative mortality" worldwide.[6] In the early 1990s, the EuroSCORE was developed, based on a multicentre European population, in an attempt to create a system more applicable to European patients.[7] More recently the American College of Cardiology/American Heart Association (ACC/AHA) task force revised their guidelines for coronary artery bypass graft surgery (CABG), including a system for prediction of outcome after isolated CABG.[8] The UK CABG Bayes model was built by the Society of Cardiothoracic Surgeons of Great Britain and Ireland (SCTS) to be used for CABG patients in the UK.[5]

Off-pump coronary artery bypass surgery (OPCAB) was introduced at the beginning of the last decade,[9] and is gaining in popularity worldwide.[10] Its use continues to be explored with case matched reports, retrospective studies, and lately prospective randomised trials.[11] Our aim in this study was to evaluate the validity and performance of the various existing risk stratification models for the rapidly growing cohort of OPCAB patients in the UK, in order to identify a suitable risk stratification system for this group of patients.

METHODS

Clinical data collection

We identified cardiothoracic surgeons from eight United Kingdom cardiac surgical centres which run established OPCAB surgery programmes using the Octopus II or III suction/stabilisation system (Medtronic Inc, Minneapolis, Minnesota, USA). They were asked to pool their clinical data on all OPCAB patients who were operated on between the start of their programmes and April 2000. The earliest operations were done in April 1996, but most of the patients (97%) were operated on between January 1998 and April 2000. Only patients undergoing first time isolated CABG for multivessel disease (more than one vessel) were included. Redo OPCAB procedures or minimally invasive direct coronary artery bypass procedures were excluded.

The participating centres were as follows: Harefield Hospital (London), Royal Brompton Hospital (London), King's College

Abbreviations: ACC/AHA, American College of Cardiology/American Heart Association; CABG, coronary artery bypass graft surgery; OPCAB, off-pump coronary artery bypass surgery; ROC, receiver operating characteristic curve; SCTS, Society of Cardiothoracic Surgeons of Great Britain and Ireland

Hospital (London), Manchester Royal Infirmary (Manchester), The Cardiothoracic Centre (Liverpool), University Hospital of Wales (Cardiff), Castle Hill Hospital (Hull), and The Royal Sussex County Hospital (Brighton).

All the clinical data were collected prospectively in line with the appended minimum dataset defined by the SCTS. The current minimum dataset, and its associated definitions, is compatible with all existing initiatives in the UK, such as UK Heart Valve Registry, the Central Cardiac Audit Database, and the British Cardiac Intervention Society database. The definitions and data fields are also compatible with evolving European initiatives and with the Society of Thoracic Surgeons, the American College of Cardiology, and the Healthcare Financing Administration in the USA.[5] Local validation of the collected data is undertaken regularly, and external validation is being done by the SCTS on a 3–5 yearly cycle. Institutional approval was obtained for the study.

Statistical analysis

Four risk scores were calculated: Parsonnet,[6] EuroSCORE,[7] ACC/AHA,[8] and the UK CABG Bayes model[5] for in-hospital mortality. These risk scores were applied to the OPCAB patient data and the observed and predicted values in clinically relevant risk groups were calculated. Their performance at predicting in-hospital mortality was then formally assessed for calibration and discrimination.

Calibration

Calibration refers to the accuracy of a score's predictions. Calibration may be assessed using the Hosmer–Lemeshow test.[12] The patients are split into five groups of roughly equal size, based on their predicted probability (to ensure the validity of the test, we only used five groups instead of the more conventional 10 groups). The predicted number of deaths in each group is compared with the number of observed deaths in each group. A significant result indicates that the observed and predicted values do not agree particularly well.

Discrimination

Discrimination refers to the ability of a score to separate sick patients from those who are less sick. Discrimination may be assessed by receiver operating characteristic curve (ROC) area.[13] The ROC area may be interpreted as the probability that a patient who died had a higher risk score than a patient who survived; thus the area under the curve is the percentage of randomly drawn pairs for which this is true. This is a fairly subjective measure and values > 0.8 usually indicate potentially useful discrimination.[14] A value of 0.5 indicates random predictions.

We could not calculate the Parsonnet score and EuroSCORE for every individual because of missing values. We used the hotdecking method[15] to impute missing values and investigated whether the calibration and discrimination values changed. Briefly, hotdecking is a multiple imputation technique that replaces missing values with values sampled at random from patients with similar characteristics. The process is repeated several times (five in this case) and the statistic of interest (the Hosmer–Lemeshow statistic and ROC area) is averaged over all values obtained. All analyses were carried out using Stata Corporation statistical software (Stata 7).[16]

RESULTS

We were able to obtain data on 2223 OPCAB patients from the eight centres. These included 571 patients from Harefield, 147 from the Royal Brompton Hospital, 164 from King's College Hospital, 577 from Manchester Royal Infirmary, 316 from the Liverpool Cardiothoracic Centre, 197 from the University Hospital of Wales, 170 from Castle Hill Hospital, and 81 from the Royal Sussex County Hospital. All data were pooled into a single database and analysed for the four risk scores. The mean (SD) age of the study patients was 62.4 (9.9) years and included 492 women (22.1%). Thirty patients (1.3%) died in hospital before discharge: 17 (0.8%) from cardiac causes, four (0.2%) from septicaemia, one (0.04%) from adult respiratory distress syndrome, and eight (0.4%) from other causes including cerebrovascular accidents, perforated bowel, mesenteric infarction, and multiple organ failure. All the surviving patients were reviewed at the outpatient clinic at six weeks postoperatively; thus the mortality figures represent the 42 day mortality.

Parsonnet

We were only able to calculate the Parsonnet score on 1515 of the 2223 patients because the two variables "body mass index" and "recently failed intervention" had missing values. We did not use the subjective variables "catastrophic states" and "other rare circumstances".

Calibration

The observed and predicted numbers of deaths in clinically relevant risk groups are presented in table 1. It is clear that the Parsonnet score considerably overestimated risk across all the risk groups, with the predicted total number of deaths (81.6) far in excess of the observed total (20). The test of calibration was highly significant ($\chi^2 = 53.1$, df = 5, $p < 0.001$), indicating very poor calibration.

Discrimination

The ROC area for the Parsonnet score was 0.74 (95% confidence interval (CI) 0.62 to 0.86), showing that this scoring system can correctly rank a pair of patients 74% of the time. As random predictions will correctly rank a pair of patients 50% of the time, the finding from this study suggest that the discriminatory power of the Parsonnet score may be limited for clinical practice in this group of patients. After imputation we were able to calculate the score for 2209 patients and found that both the calibration ($\chi^2 = 77.6$, df = 5, $p < 0.001$) and discrimination (0.69, 95% CI 0.59 to 0.79) deteriorated.

EuroSCORE

We were able to calculate the EuroSCORE in 1907 patients. We did not have the factor "pulmonary hypertension", so the effect of this was not incorporated into the score.

Calibration

The observed and predicted numbers of deaths in clinically relevant risk groups are shown in table 1. EuroSCORE appears reasonably well calibrated for the highest risk group but is not so well calibrated for the other groups. The predicted total number of deaths (49.6) is nearly double the observed total (26). The test of calibration was significant ($\chi^2 = 13.8$, df = 5, $p = 0.008$). The calibration was better than for the Parsonnet score though it was still poor.

Discrimination

The ROC area for the EuroSCORE was very similar to that of the Parsonnet score (0.75, 95% CI 0.64 to 0.85).

After imputation we were able to calculate the score for 2221 patients and found that the calibration deteriorated ($\chi^2 = 16.2$, df = 5, $p = 0.006$), while the discrimination improved slightly (0.77, 95% CI 0.67 to 0.86).

ACC/AHA

We were able to calculate the ACC/AHA score for all the 2223 patients.

Calibration

The observed and predicted numbers of deaths in clinically relevant risk groups are shown in table 1. It is clear that there was good agreement between the observed and predicted

Table 1 Comparison of observed and predicted mortality using the four risk scores in clinically relevant risk groups

	Parsonnet score			EuroSCORE			ACC/AHA			UK CABG Bayes		
Risk group	n	Obs (%)	Exp (%)	n	Obs (%)	Exp (%)	n	Obs (%)	Exp (%)	n	Obs (%)	Exp (%)
0–2.49%	472	2 (0.4)	2.8 (0.6)	1061	6 (0.6)	10.1 (0.9)	2016	17 (0.8)	18.7 (0.9)	1798	13 (0.7)	15.1 (0.8)
2.5–4.99%	371	1 (0.3)	12.0 (3.2)	478	5 (1.0)	16.4 (3.4)	130	5 (3.8)	4.1 (3.1)	246	4 (1.6)	8.5 (3.4)
5.0–9.99%	348	5 (1.4)	22.3 (6.4)	344	12 (3.5)	20.6 (6.0)	68	7 (10.3)	4.3 (6.3)	122	7 (5.7)	8.1 (6.7)
10–19.99%	294	11 (3.7)	37.7 (12.8)	24	3 (12.5)	2.5 (10.4)	9	1 (11.1)	1.1 (12.3)	40	4 (10.0)	5.6 (14.0)
≥20%	30	1 (3.3)	6.8 (22.8)	0			0			17	2 (11.8)	5.1 (29.7)
Total	1515	20 (1.3)	81.6 (5.4)	1907	26 (1.4)	49.6 (2.6)	2223	30 (1.3)	28.2 (1.3)	2223	30 (1.3)	42.4 (1.9)

ACC/AHA, American College of Cardiology/American Heart Association model; Exp, expected; Obs, observed; UK CABG, Society of Cardiothoracic Surgeons of Great Britain and Ireland coronary artery bypass graft surgery model.

values. However, the range of predictions was fairly limited (the largest prediction was only 15.7%). The test of calibration was non-significant ($\chi^2 = 2.2$, df = 5, p = 0.71), suggesting that the ACC/AHA score is well calibrated.

Discrimination

The ROC area for the ACC/AHA score was 0.75 (95% CI 0.64 to 0.85), similar to the Parsonnet and EuroSCORE scores.

UK Bayes model

We were able to calculate the UK Bayes model score for all the 2223 patients. Some patients had missing predictor values but the model allows for this by effectively imputing average predictor values in place of the missing values.

Calibration

The observed and predicted numbers of deaths in clinically relevant risk groups are shown in table 1. It is clear that there was a reasonably good agreement between the observed and predicted values, apart from the highest risk group (which contained only 17 patients). The predicted total number of deaths (42.4) exceeded the observed total (30), suggesting that the score slightly overestimates the risk of mortality. However, the test of calibration was non-significant ($\chi^2 = 6.1$, df = 5, p = 0.3), indicating that the UK CABG Bayes model is reasonably well calibrated.

Discrimination

The ROC area for the UK Bayes model score was 0.81 (95% CI 0.73 to 0.88). Therefore, the UK Bayes model had the best discriminatory power of the four risk models for these data. We also noted that the lower bound of the confidence interval (0.73) was almost as large as the ROC areas achieved with the other scores.

DISCUSSION

Our study suggests that the ACC/AHA and UK Bayes models may be suitable risk adjustment models for this group of OPCAB patients, as both predict the risk of mortality reasonably accurately. Of the two, however, we feel that the UK Bayes model is superior, as it provides better discrimination. The range of predictions provided by the ACC/AHA score is limited and perhaps clinically unrealistic, because most of the patients are assigned a risk of < 2.5% and the highest risk is only 15.7%. This is reflected in the relatively low ROC area. In contrast, the UK Bayes model makes a wider range of predictions yet still remains accurate.

We were able to calculate a UK Bayes model score for all patients because this method assigns average scores when the value of a particular predictor is missing for a particular patient.[17 18] The other scores do not have this option. "Pulmonary hypertension", required by the EuroSCORE, is not readily available on all patients undergoing coronary surgery in the UK, with the result that this scoring system is not readily applicable in this country. The absence of this variable may have had a detrimental effect of the performance of EuroSCORE in our study. The subjective variables "catastrophic states" and "other rare circumstances" can have a major effect on the calculation of the Parsonnet score, and it has been suggested that they should not be used.[5]

Risk stratification plays a vital role in the cardiac surgical practice throughout the world. Hospitals, universities, institutions, and health authorities have realised the importance of assessing the clinical outcomes of cardiac surgery in an objective risk adjusted manner, as this allows valid and realistic comparisons to be made between countries, regions, hospitals, and even individual surgeons in both a longitudinal and a cross sectional fashion.[5] Furthermore, risk models can detect and quantify differences and changes in the risk profiles of patients presenting for cardiac surgery.[3] By relating risk

factors to surgical outcomes, the risk models provide an important tool to assess the effect of the changes in surgical techniques or managements and help plan for the optimal use of available resources.[4] Most importantly, it allows an objective assessment of the surgeon's performance and gives the opportunity to the patient to give well informed consent.[5]

Over the last decade, many risk stratification systems have been developed using logistic regression and Bayes modelling techniques. Statisticians also developed their tools to assess the performance of those systems for precisely predicting the observed outcomes. All the risk scores were developed on patients undergoing only, or mainly, on-pump cardiac procedures. The use of cardiopulmonary bypass has been found to be an independent risk factor for in-hospital mortality, so we might expect the risk of mortality to be overestimated in the OPCAB patients using those scores.[19] This is generally the case, although some of the scores have also been shown to overestimate mortality in patients treated with cardiopulmonary bypass.[20]

Patients presenting for cardiac surgery are a heterogeneous group differing greatly in their risk profiles, the effect of those risk factors on the outcome, the hospitals where they are operated on,[21] the surgeons who operate, and even the type of surgery (whether valve or coronary[22] and, more recently, whether on-pump or off-pump). It is not surprising to find that even the intraoperative physiological variables can affect in-hospital mortality.[23] Our findings support the concept that one single risk score cannot predict mortality precisely in a heterogeneous group of patients, and we suggest that risk stratification systems should be single procedure specific[24] and perhaps geographical location specific.

Conclusions

Our study suggests that among the currently available risk scores, the UK Bayes model is the best risk stratification model for application on OPCAB patients in the UK. However, larger studies are required to confirm these results or create a new specific risk stratification system for this growing group of patients.

ACKNOWLEDGEMENTS

The statistical analysis was supported by a grant from Garfield Weston Trust. We would like to acknowledge the assistance we received from Janet Deane (Liverpool), Joe Omigie (Kings College), Suzanne Chaisty (Manchester), and Nilanjan Chaudhuri (Hull) in the data collection process. Presented at the annual meeting of The Society of Cardiothoracic Surgeons of Great Britain and Ireland (17–20 March 2002, Bournemouth, UK), and at the annual meeting of the International Society of Minimally Invasive Cardiac Surgery (20–23 June 2002, New York, USA).

Authors' affiliations

S Al-Ruzzeh, G Asimakopoulos, K Taylor, M Amrani, The National Heart and Lung Institute, Harefield and Hammersmith Hospitals, London, UK
G Ambler, R Omar, Department of Statistical Science, University College London, London, UK
B Fabri, M Pullan, Cardiothoracic Centre, Liverpool, UK
A El-Gamel, King's College, London, UK
R Hasan, D Keenan, Manchester Royal Infirmary, Manchester, UK
Z Zamvar, University College of Wales, Cardiff, UK
S Griffin, A Cale, M Cowen, Castle Hill Hospital, Hull, UK
A De Souza, Royal Brompton Hospital, London, UK
U Trivedi, Royal Sussex County Hospital, Brighton, UK

REFERENCES

1 **Hammermeister K**. Risk, predicting outcomes and improving care. *Circulation* 1995;91:899–900.
2 **Parsonnet V**, Bernstein A, Gera M. Clinical usefulness of risk-stratified outcome analysis in cardiac surgery in New Jersey. *Ann Thorac Surg* 1996;**61**:S8–11.
3 **Nashef S**, Roques F, Michael P, *et al*. Coronary surgery in Europe: comparison of the national subsets of the European System for Cardiac Operative Risk Evaluation database. *Eur J Cardiothorac Surg* 2000;**17**:396–9.
4 **Smith P**, Smith L, Muhlbaier L. Risk stratification for adverse economic outcomes in cardiac surgery. *Ann Thorac Surg* 1997;**64**:S61–3.
5 **Anonymous**. The Society of Cardiothoracic Surgeons of Great Britain and Ireland. National adult cardiac surgical database report. Final draft. London: SCTS, 1998.
6 **Parsonnet V**, Dean D, Bernstein A. A method of uniform stratification of risk for evaluating the results of surgery in acquired heart disease. *Circulation* 1989;**79**:I3–12.
7 **Nashef S**, Roques F, Michael P, *et al*. European system for cardiac operative risk evaluation (EuroSCORE). *Eur J Cardiothorac Surg* 1999;**16**:9–13.
8 **Eagle K**, Guyton R, Davidoff R, *et al*. ACC/AHA Guidelines for coronary artery bypass graft surgery: executive summary and recommendations. *Circulation* 1999;**100**:1464–80.
9 **Benetti F**, Naselli G, Wood M, *et al*. Direct myocardial revascularization without extracorporeal circulation. Experience in 700 patients. *Chest* 1991;**100**:312–16.
10 **Yacoub M**. Off-pump coronary bypass surgery in search of an identity. *Circulation* 2001;**104**:1743–5.
11 **Van Dijk D**, Nierich A, Jansen E, *et al*. Early outcome after off-pump versus on-pump coronary bypass surgery: results from a randomised study. *Circulation* 2001;**104**:1761–6.
12 **Hanley J**, McNeil B. The meaning and use of the area under a receiver-operating-characteristic curve. *Radiology* 1982;**143**:29–36.
13 **Hosmer D**, Lemeshow S. *Applied logistic regression*. New York: John Wiley, 1989.
14 **Harrell F**. *Regression model strategies*. New York: Springer Verlag, 2001.
15 **Schafer J**. *Analysis of incomplete multivariate data by simulation*. London: Chapman and Hall, 1995.
16 **StataCorp**. *Stata statistical software: release 7*. College Station, Texas: Stata Corporation, 2001.
17 **Marshall G**, Shroyer L, Grover F, *et al*. Bayesian-logit model for risk assessment in coronary artery bypass grafting. *Ann Thorac Surg* 1994;**57**:1492–500.
18 **Edwards F**, Albus R, Zajtchuk R, *et al*. A quality assurance model of operative mortality in coronary artery surgery. *Ann Thorac Surg* 1989;**47**:646–9.
19 **Calafiore A**, Di Mauro M, Contini M, *et al*. Myocardial revascularization with and without cardiopulmonary bypass in multivessel disease: impact of the strategy on early outcome. *Ann Thorac Surg* 2001;**72**:456–62.
20 **Bridgewater B**, Neve H, Moat N, *et al*. Predicting operative risk for coronary artery surgery in the United Kingdom: a comparison of various risk prediction algorithms. *Heart* 1998;**79**:350–5.
21 **Clark R**. Outcome as a function of annual coronary artery bypass graft volume. *Ann Thorac Surg* 1996;**61**:21–6.
22 **Jamieson W**, Edwards F, Schwartz M, *et al*. Risk stratification for cardiac valve replacement. National cardiac surgery database. *Ann Thorac Surg* 1999;**67**:943–51.
23 **Hill S**, van Wermeskerken G, Lardenoye J, *et al*. Intraoperative physiologic variables and outcome in cardiac surgery: part I. In-hospital mortality. *Ann Thorac Surg* 2000;**69**:1070–6.
24 **Chertow G**, Lazarus M, Christiansen C, *et al*. Preoperative renal risk stratification. *Circulation* 1997;**95**:878–84.