An evaluation of existing risk stratification models as a tool for comparison of surgical performances for coronary artery bypass grafting between institutions


*Imperial College School of Medicine at Hammersmith Hospital, London, UK
bImperial College School of Medicine at Harefield Hospital, London, UK
cDepartment of Statistical Science, University College, London, UK

Received 13 November 2002; received in revised form 22 February 2003; accepted 12 March 2003

Abstract

Objective: Risk stratification systems are used in cardiac surgery to estimate mortality risk for individual patients and to compare surgical performance between institutions or surgeons. This study investigates the suitability of six existing risk stratification systems for these purposes.

Methods: Data on 5471 patients who underwent isolated coronary artery bypass grafting at two UK cardiac centres between 1993 and 1999 were extracted from a prospective computerised clinical database. Of these patients, 184 (3.3%) died in hospital. In-hospital mortality risk scores were calculated for each patient using the Parsonnet score, the EuroSCORE, the ACC/AHA score and three UK Bayes models (old, new complex and new simple). The accuracy for predicting mortality at an institutional level was assessed by comparing total observed and predicted mortality. The accuracy of the risk scores for predicting mortality for a patient was assessed by the Hosmer-Lemeshow test. The receiver operating characteristic (ROC) curve was used to evaluate how well a system ranks the patient with respect to their risk of mortality and can be useful for patient management.

Results: Both EuroSCORE and the simple Bayes model were reasonably accurate at predicting overall mortality. However predictive accuracy at the patient level was poor for all systems, although EuroSCORE was accurate for low to medium risk patients. Discrimination was fair with the following ROC areas: Parsonnet 0.73, EuroSCORE 0.76, ACC/AHA system 0.76, old Bayes 0.77, complex Bayes 0.76, simple Bayes 0.76. Conclusions: This study suggests that two of the scores may be useful in comparing institutions. None of the risk scores provide accurate risk estimates for individual patients in the two hospitals studied although EuroSCORE may have some utility for certain patients. All six systems perform moderately at ranking the patients and so may be useful for patient management. More results are needed from other institutions to confirm that the EuroSCORE and the simple Bayes model are suitable for institutional risk-adjusted comparisons.

Keywords: Risk stratification; Mortality; Cardiac surgery; Clinical aims

1. Introduction

Risk stratification models are increasingly used in cardiac surgery to investigate patient outcomes in relation to patient and pre-operative disease characteristics. Since the mid-1980s, many risk stratification systems have been developed [1–10]. These models estimate coefficients for each risk factor of mortality which are often translated to risk scores. The scores assigned to each risk factor are then added to calculate the overall risk score of mortality for a patient and to construct clinical risk groups. Reference to these groups is made for various purposes, for example to adjust clinical decisions to individual patients’ circumstances and prognosis, to compare surgical performances and for patient counselling. It is desirable that these models should be useful for outcome prediction at different surgical centres, both at the overall institutional level as well as the
patient level, and, if possible, in different countries. Following formulation, most risk models are initially validated using a patient sample from the same institutions as those which provided the original patient sample used for model formulation. In general, models predict outcome more accurately when used in the original setting than when applied to other patient populations [1]. The more rigorous form of validation is to evaluate the model on new data involving patients undergoing surgery at a subsequent time period and possibly at a different centre. The clinical aim of the model (for example, to compare institutions, or for patient advice or management) should be taken into account when performing this task.

The Parsonnet system was developed in the US and was one of the first systems for predicting risk in cardiac surgery [2]. It is widely used in the UK although it has been criticised for including subjective variables [11]. The EuroSCORE is similar in concept to the Parsonnet score and was developed using data from 128 European cardiac surgical centres [8]. Recently the American College of Cardiology/American Heart Association Task Force revised their guidelines for Coronary Artery Bypass Graft Surgery, including a system for prediction of outcome after isolated coronary artery bypass grafting (CABG) surgery [10]. All three of these risk scores use simple, additive scoring systems.

The Society of Cardiothoracic Surgeons of Great Britain and Ireland (SCTS) have proposed the use of a Bayes Model for CABG patients in the UK [12]. More recently the Society developed a new ‘complex Bayes model’ and a new ‘simple Bayes model’ [11]. The nine factor complex Bayes model is a subset of the old model while the five factor simple Bayes model is derived from the complex model. These models are designed to automatically handle missing values for risk factors by effectively assigning an average risk to the missing factor or category. Hence risk scores can be calculated for all patients. The risk factors included in the six models described above are outlined in Table 1.

This study evaluates the six risk stratification models using patients undergoing isolated CABG at two British cardiac surgical centres. The objective of the study is to investigate the suitability of these systems for comparison of risk stratified surgical performance at an institutional level, patient counselling and treatment management.

2. Patients and methods

This study was conducted in collaboration between the Cardiothoracic Departments at Hammersmith and Harefield Hospitals, London, UK. Data were extracted from the computerised database in each department. All the clinical data were collected prospectively in line with the appended Minimum Dataset (MDS) defined by the SCTS. The current MDS, and its associated definitions, is compatible with all existing initiatives in the UK such as the 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 the Society of Thoracic Surgeons (USA), American College of Cardiology and the Healthcare Financing Administration (HCFA) in the United States [11]. Local validation of the collected data is performed regularly and external validation is being performed by the Society on 3–5-yearly cycle.

Scores for the different risk stratification models were calculated retrospectively using data of the 5471 patients who underwent isolated CABG at the participating hospitals between January 1993 and December 1999. The 13 patients who underwent salvage procedure were not included. Salvage patients were excluded because of the strongly subjective element in applying the Parsonnet score in this special group of patients. Patients undergoing other concomitant procedures, such as valve replacement, were also excluded from this study. All procedures were carried out with cardiopulmonary bypass. The outcome measurement considered in the analysis was hospital mortality, defined as death occurring before hospital discharge. Predicted hospital mortality was calculated separately for the Parsonnet system, ACC/AHA system, EuroSCORE, old UK Bayes, complex Bayes and simple Bayes systems, using the risk scoring systems suggested in the original publications of these risk models.

2.1. Statistical analysis

Calibration signifies the degree of correspondence between the actual mortality and the mortality predicted by each risk model. We evaluated the predictive accuracy of the risk stratification systems for both institutional comparisons and patient evaluation. For the former we considered the agreement between the total observed mortality and the total predicted mortality. A 95% reference interval was constructed around the total predicted mortality and we considered whether the total observed mortality lay within this. The total predicted mortality is the sum of all the individual patient predictions (when expressed as probabilities) and the interval takes into account the uncertainty in this total (since the predictions are probabilities, not certainties).

To evaluate the risk stratification systems at the patient level (that is, whether it can be used to predict mortality for an individual patient) we used the Hosmer-Lemeshow (H-L) test [13]. This test evaluates the correspondence between observed and predicted mortality within a number of risk groups. The smaller the value of the H-L test statistic the better the calibration. A P-value of <0.05 indicates that statistically the model is significantly a bad fit to the data and is not predicting the risk of mortality accurately. To carry out the H-L test, the patients were split into six clinical risk groups, based on preoperative predicted mortality (<1, 1–2, 2–3, 3–5, 5–10 and >10%).
Discrimination is the ability of the system to discriminate between patients who will die in the hospital following surgery and patients who will survive. Discrimination was assessed using the receiver-operating-characteristic (ROC) curve area. Values of 0.5 indicate that model cannot discriminate better than chance and 1 indicates perfect discrimination.

All analysis was carried out using the statistical software Stata 7 (Stata Corporation, USA) [14].

### 3. Results

Comparison of the pre-operative risk factors of patients between the two cardiac centres showed that the patients were similar in terms of age, gender, operative priority, level of hypertension and prevalence of respiratory and cerebrovascular diseases, previous myocardial infarction and diabetes. There were some differences for a small number of pre-operative conditions of the patients between the two hospitals. These were renal failure (higher in Hammersmith, 3.2 versus 0.3%), poor ejection fraction (higher in Hammersmith, 11.8 versus 5.2%), Left Main Stem Disease >50% (higher in Harefield, 9.4 versus 5.9%), use of IABP (higher in Harefield, 3.3 versus 1.0%). The data from the two centres were combined to evaluate the six models. Of the 5471 patients included in the study, 184 (3.3%) died in hospital. The mean age of the patients was 60.6 years (standard deviation: 9). There were 4705 (86%) male and 766 (14%) female patients.

Data allowing the calculation of risk scores for the Parsonnet, EuroSCORE and the ACC/AHA models were not available for all patients. We were only able to calculate the Parsonnet score for 4439 patients mainly because the two risk factors body mass index and recently failed intervention used in this model had missing values. As suggested by the SCTS [11] we did not use the subjective risk factors catastrophic states and other rare circumstances that were included in the original Parsonnet model. We were able to calculate the EuroSCORE for 4654 patients. We did not have pulmonary hypertension so the effect of this was not incorporated into the calculation of the score. We

### Table 1

<table>
<thead>
<tr>
<th>Risk factor</th>
<th>Parsonnet</th>
<th>ACC/AHA</th>
<th>Euro-SCORE</th>
<th>SCTS-old</th>
<th>SCTS-simple</th>
<th>SCTS-complex</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sex</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Obesity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chronic pulmonary disease</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Extracardiac arteriopathy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hypertension</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diabetes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ejection fraction/LV dysfunction</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neurological dysfunction disease</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Previous cardiac surgery</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Preoperative IABP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recent thrombolysis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Renal problems</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Active endocarditis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Critical pre-operative state</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Catastrophic statesa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rare circumstancesa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Left ventricular aneurysm</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unstable angina</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recent cardiac infarct</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cardiac procedure other than CABG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Peripheral vascular disease</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Surgical priority</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Left main stem disease</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Postinfarct septal rupture</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recently failed intervention</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pulmonary hypertension</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Surgery on thoracic aorta</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Body surface area</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relevant to valve surgery</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mitral – systolic PA pressure</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aortic – AV pressure gradient</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CABG at time of valve surgery</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a Subjective risk factors present in Parsonnet score – often not used in practice.
calculated the ACC/AHA score for 4753 patients, as information on operative priority required by this model was occasionally missing. We investigated the available pre-operative characteristics of patients who had some risk factors missing. They were similar to those of the other patients in the study.

3.1. Calibration

3.1.1. Overall mortality

Table 2 shows the total number of observed deaths alongside the total number of predicted deaths and the corresponding 95% reference interval. EuroSCORE and the simple Bayes model predict overall mortality reasonably well with the observed totals lying inside the reference intervals. The observed total lies right on the edge of the reference interval from the complex Bayes model. The other three models are not accurate with the ACC/AHA score grossly underestimating the risk of mortality and the Parsonnet model overestimating mortality.

3.1.2. Mortality in risk groups

Fig. 1 presents the distribution of patients across the six clinical risk groups for each model. Fig. 2 shows a comparison of the observed deaths with those predicted by the six cardiac risk models. Parsonnet overestimates the risk of mortality across all risk groups. EuroSCORE predicts mortality reasonably accurately for the low to medium risk patients but performs badly for the high risk patients. The ACC/AHA model consistently underestimates the risk of mortality. The old Bayes model under predicts mortality for the low to medium risk group of patients. The two new Bayes models under predict mortality for the medium risk patients and over predict mortality for the high risk group of patients.

The numbers of patients and the results from the H-L test for the six models are shown in Table 3. These results and Fig. 2 suggest that all models show poor calibration at the patient level (P = 0.004 for the simple Bayes model and <0.001 for the rest). The values of the H-L test statistic suggest that the ACC/AHA score produces the most inaccurate predictions of mortality for these patients, followed by the Parsonnet score and the old Bayes model. The best predictions are made by the simple Bayes model, followed by the EuroSCORE.

4. Discussion

Risk stratification in cardiac surgery has been an area of increasing importance in recent years. Health authorities, hospitals and individuals, such as medical practitioners and patients, are placing importance in the objective risk-adjusted prediction of mortality after cardiac surgery. Risk stratification models aim for a more objective comparison of surgical performance between institutions or individual surgeons. They can detect and quantify differences and changes in the risk profiles of patients presented for cardiac surgery. Furthermore, risk prediction allows a more objective assessment of the indication for surgery in individual patients by facilitating more accurate balancing of potential risks and benefits [10,11].

Risk stratification models have been criticised for reduced applicability when used in different patient populations to the ones they were formulated on. Models developed in the US, for instance, may not predict satisfactorily clinical outcome in European patient populations. Developing a statistical model that predicts risk of death after cardiac surgery with a high degree of accuracy, is
desirable for various reasons. It facilitates meaningful informed consent of the patient and contributes to the decision whether the potential benefit of surgery for a particular individual outweighs the potential risk. This knowledge assists in defining the indication for surgery, improves communication between the physician and the patient and ultimately improves patient care. At a collective level, analysis of patient outcome in relation to predicted risk allows individual surgeons and institutions to evaluate their results and compare them to others. Meaningful audit allows evaluation of outcome and is likely to protect clinicians from medico-legal litigation. Furthermore, by estimating which patient will benefit from surgery, risk prediction contributes to better allocation of resources.

This study evaluates six existing risk stratification systems for CABG surgery with respect to three specific clinical aims: (a) to compare overall institutional performance; (b) to provide patient advice; and (c) to manage
We use data from 5471 patients operated on between 1993 and 1999 at two British cardiac surgical centres.

The Parsonnet, EuroSCORE, and ACC/AHA models present simple risk scores that can be used easily for patient consultation by facilitating rapid prediction of risk. The Bayes models on the other hand, require more complex calculation of risk scores, usually performed by computers.

Several studies have validated the Parsonnet model using independent data [3–6,11,15]. The Parsonnet model has been shown to over-estimate mortality [3,5,6,15] and this is the case in our study. This may be attributed to the methodology used to develop the score and this has been criticised [16,17]. The EuroSCORE predicts overall mortality reasonably well in our study and this has also been demonstrated in other studies [9,18]. The score also made good predictions in our study at the patient level for low to medium-risk patients (97% of the sample). We were unable to include pulmonary hypertension in the risk score calculation since this measurement is rarely available for patients undergoing isolated CABG in the UK. This could be partially responsible for this model predicting poorly for the high risk group of patients.

The ACC/AHA model predicts risk of mortality as well as stroke and mediastinitis, and the small number of variables make it considerably simpler than the EuroSCORE and Parsonnet models. However our study suggests that it grossly under-estimates mortality. There are no published studies to date that that have carried out a validation of this scoring system using external data. The Bayes models have been validated using patients from the same centres, undergoing surgery at similar time periods as that of the patients used in the model development process and have been shown to perform reasonably well [11]. However no other published studies have carried a validation of these models on the basis of completely independent data. All the Bayes systems underestimated mortality for low to medium risk patients in our study although the simple Bayes model predicted overall mortality reasonably well.

At the patient level, the calibration results (Fig. 2 and Table 3) suggest that risk models should be used with caution when informing a patient about their chance of dying in hospital following surgery. On the basis of the findings from this study, none of the risk models can be completely relied on to give accurate information to patients although EuroSCORE is seems accurate for all but high-risk patients.

Five of the models achieved similar ROC areas in our study (0.76–0.77). The ROC area for Parsonnet was considerably lower (0.73). These ROC areas are comparable with those of the UK National Adult Cardiac Surgical Database [11], 0.71 for Parsonnet, 0.75 for EuroSCORE, 0.74 for simple Bayes and 0.75 for complex Bayes, while no externally calculated ROC area has been published previously for the ACC/AHA system. When assessed at different single centres, the ROC area for Parsonnet system varied from 0.65 to 0.85 [3–6,9], while the EuroSCORE produced ROC areas of 0.75–0.78 in previous studies [9,18].

In this study we have concentrated on short-term mortality. However this may not by itself be an adequate indicator of quality of care or resource use. Morbidity being more common than mortality may be more informative and can be measured in terms of post-operative complications and length of stay in hospital. Long-term mortality, which may be a more useful outcome, is rarely assessed probably...
because of the difficulty in following patients over long periods of time [19]. However in countries, such as the UK, with good death registration systems this is achievable and should be a priority for future research in risk modelling.

We have considered the performance of the risk scores but have not commented on the methodology used to develop them. This was beyond the scope of this paper although we do discuss the issues elsewhere [20]. Also we have not commented on the ability of the risk systems to be used for comparing surgeons. This is because we do not have information regarding the surgeons involved. However we believe that if a model is well-calibrated at the patient level then it should be suitable for this purpose. However, more research is required to demonstrate this.

In summary, this study evaluates six risk stratification models in cardiac surgical patients on a completely new patient sample from cardiac centres in the UK. Good calibration at the institutional level is essential for performing risk-stratified comparisons between institutions. In our study two of the models, EuroSCORE and simple Bayes, predicted the overall level of mortality in these data reasonably well (Table 2). These results suggest that these scores may be appropriate for producing case-mix adjusted league tables to assess institutional performances. However we would need to see similar findings from other institutions before drawing a firm conclusion. None of the six models excelled at producing good mortality estimates for an individual patient. However EuroSCORE proved to be accurate for low to medium risk patients which is the majority of patients. We note that the SCTS [11] comment on the difficulty of producing accurate predictions for high-risk patients. In term of discrimination, at least five of the models showed a moderate ability to differentiate between low and high risk patients which may help in making treatment decisions and managing surgical lists. The perfect risk stratification system is still eluding us but some of these risk models, used with care, may have some value for institutional comparisons and patient counselling.

Acknowledgements

The statistical analysis for this project was supported by a grant from the Garfield Weston Trust. We thank the reviewers for helpful comments.

References


Appendix A. Conference discussion

Dr F. Grover (Denver, CO): A very nice, succinct presentation. I have enjoyed working with Ken Taylor over the years in this area. I have had the privilege over the past several years of chairing the Society of Thoracic Surgeons National Database Committee and Work Force, and in the United
States we have had a decade or so to consider some of these issues because of being somewhat pushed into it by HCFA, our Social Security or Medicare system back in the mid 1980s.

The C index of 0.75 or 0.76 or the ROC curve leaves about 25% of the predictability to either the way that a hospital or the physicians or the whole team takes care of the patients, or by chance, and that is probably a reasonable amount of difference to expect according to the variability and care.

I think the main value for these systems is our ability as surgeons to look at our own data and compare our data to those of others, or at least to the aggregate, nationally or in Europe, whatever the geographic area is, and then you kind of see if you are doing really well or maybe not so well and areas that you can improve. When you get into comparing hospital to hospital, as you mentioned, that is always a more ticklish issue.

I think certainly the STS, which uses a logistic regression model and has had the chance to mature over the years with enough patients in the system, it now has roughly 25 variables that are predictive of operative mortality, and our goodness of fit or calibration, so to speak, is actually very good, being the least accurate at the very, very high risk because there are fewer patients in the very high risk.

We found that if we don’t do this ourselves, somehow to kind of look at ourselves and occasionally identify hospitals, the government will do it for us, and I guess the main gist I would like to say in this discussion is that I think you need to keep pressing ahead for better systems, and I would urge you to be very proactive as a professional organisation, taking charge of the kind of data, have the surgeons collect the data rather than have the government do it, or else they will do it for you and it won’t be done nearly as well.

Mr B. Keogh (Birmingham, UK): Thank you very much for a very interesting presentation, which I think illustrates some of the concerns that many of us have over the overuse and over-confidence in risk stratification. But there is another message I think from your paper, and that is that risk stratification models are designed on large population groups and are designed to predict variance and performance within those population groups.

So, for example, the EuroSCORE is designed on a huge European database and the Bayes models for the UK which you allude to are derived from a large number of UK patients, and you would expect the predictive value of those to vary from hospital to hospital within those groups; they will overscore in 50% of hospitals and underscore in 50% of hospitals.

So I think you must expect this kind of variance, and to attempting to validate a national model using only two hospitals is less a validation of the model, I think, and more a study of the variance of performance within those hospitals. But having said that, I think your point is very well taken, that these models should simply be used as pointers to look elsewhere, and I think Fred Grover’s point, that we should strive for more accurate models, is also extremely important.

Dr Asimakopoulos: Probably a quick comment on that to emphasise my main point, as Mr Keogh said. One of the aims of the study is that when next time the UK government or any government publishes a list of hospitals based on mortality, it would be inappropriate, at least based on the validation of the study in our two hospitals, to use any of the existing systems in order to rank the hospitals based on mortality. A better more accurate system will have to be developed first.