Mortality risk prediction in coronary surgery: a locally developed model outperforms external risk models
Pedro E. Antunes, Luís Eugénio, J. Ferrão de Oliveira and Manuel J. Antunes
Interact CardioVasc Thorac Surg 2007;6:437-441; originally published online Apr 6, 2007;
DOI: 10.1510/icvts.2007.152017

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://icvts.ctsnetjournals.org/cgi/content/full/6/4/437
Mortality risk prediction in coronary surgery: a locally developed model outperforms external risk models

Pedro E. Antunes, Luís Eugénio, J. Ferrão de Oliveira, Manuel J. Antunes*

Cardiothoracic Surgery, University Hospital, Coimbra, Portugal

Received 11 January 2007; received in revised form 7 March 2007; accepted 7 March 2007

Abstract

This study aimed at assessing the performance of three external risk-adjusted models — logistic EuroSCORE, Parsonnet score and Ontario Province Risk (OPR) score — in predicting in-hospital mortality in patients submitted to coronary artery bypass graft (CABG) and to develop a local risk-score model. Data on 4567 patients who underwent isolated CABG (1992–2001) were extracted from our clinical database. Hospital mortality was 0.96% (44 patients). For the three external systems, observed and predicted mortalities were compared, and discrimination and calibration were assessed. A local risk model was developed and validated by means of logistic regression and bootstrap analysis. The EuroSCORE predicted a mortality of 2.34% (P < 0.001 vs. observed), the Parsonnet 4.43% (P < 0.0001) and the OPR 1.66% (P < 0.005). All models overestimated mortality significantly in almost all tertile risk groups. The areas under the ROC curve (AUC) for EuroSCORE, Parsonnet and OPR were 0.754, 0.664 and 0.683, respectively. The local model exhibited good calibration and discrimination AUC, 0.752. In conclusion, the three risk-score systems analyzed do not accurately predict in-hospital mortality in our coronary surgery patients; hence their use for risk prediction may not be appropriate in our population. We developed a risk-prediction model that can be used as an instrument to provide accurate information about the risk of in-hospital mortality in our patient population.

© 2007 Published by European Association for Cardio-Thoracic Surgery. All rights reserved.

Keywords: Coronary artery bypass surgery; Mortality predictive models; Risk-adjusted mortality

1. Introduction

Decision-making, preoperative patient education and consent, and quality control are some important areas where risk prediction models play an important role in current cardiac surgical practice [1]. These targets require scoring systems with good performance on three aspects: discrimination, calibration and stability over a wide spectrum of risk.

We had previously sensed that the most commonly used risk-score systems did not accurately predict mortality in our patient population, but had yet not been able to analyze and quantify the discrepancy. Hence, our purpose was first, to assess the performance of three risk-adjusted predictive models — the EuroSCORE [2], the Parsonnet score [3] and the Ontario Province Risk score [4] — in predicting in-hospital mortality in our patients submitted to CABG. Secondly, to develop and validate a risk model for in-hospital mortality with the aim to provide information to clinicians and patients about the risk in our patient population anticipating CABG.

2. Materials and methods

2.1. Data

The present study includes 4567 patients submitted to isolated CABG at our institution in a 10-year period, from January 1992 through December 2001. Pre-operative, operative and in-hospital outcome data were prospectively collected.

There were 4030 men (88.2%) and 537 women and the mean age was 60.6 ± 9.2 years. All operations were performed under hypothermic ventricular fibrillation, without cardioplegia, or empty beating heart, a technique described in detail in previous reports [5, 6]. The mean number of grafts per patient was 2.8 ± 0.8 and mean cardiopulmonary bypass time was 63.3 ± 22.9 min. The endpoint of the study was in-hospital mortality, defined as death during hospital stay, unlimited in time. All survivors were discharged to their home. The overall observed in-hospital mortality was 44 patients (0.96%). The interval between surgery and death ranged from 1 to 127 days, and six deaths (13.6%) occurred beyond 30 days.

2.2. Analysis

2.2.1. Performance of external risk models

Three risk scores were calculated retrospectively: the logistic EuroSCORE, the Parsonnet score and the Ontario Province Risk (OPR) score [2–4].
Definitions of four of the risk factors in our database differed from those of the EuroSCORE. However, some adjustments or approximate assumptions were made to enable the analysis (Table 1), a methodology previously used by others [7, 8]. We did not have data on pulmonary hypertension and critical pre-operative state, hence the effect of these risk factors was not incorporated into the calculation. We obtained a good definition match between our variables and the Parsonnet and OPR risk factors but, as suggested by others [9], we did not use the subjective risk factors catastrophic states and other rare circumstances that were included in the original Parsonnet model.

A logistic regression analysis of in-hospital mortality on the resulting scores was performed, which enabled the measurement of both the discrimination and the calibration of each of these scores on our population. Performance of the models was also assessed by comparing the observed and expected mortality in tertiles of risk. The percent difference between the predicted and observed hospital mortality was calculated using the following formula: \((\text{predicted deaths} - \text{observed deaths}) \times 100 / \text{predicted deaths}\). The Fisher exact test was used for the contingency tables.

### 2.2.2. Local risk prediction model for in-hospital mortality

More than 50 pre-operative patient variables were available from the database, of which 21 potential risk factors were chosen, identified from clinical knowledge and previous research [10] (Appendix A). The entire database was initially used to develop the predictive logistic model. Survivors and non-survivors were initially compared by univariate analysis performed with the unpaired Student’s t-test or the Mann–Whitney test for numeric variables, and the \(\chi^2\) test or the Fisher exact test for categorical variables. Variables with a \(P<0.2\) at univariate analysis were used as independent variables in a forward stepwise logistic regression analysis with in-hospital mortality as the binary dependent variable. Because of the relatively small effective sample size (44 deaths), a \(P<0.1\) was selected for variable retention in the final regression model. A bootstrap analysis was used in combination with the logistic regression analysis to select the final set of risk factors included in the model. In the bootstrap procedure, 200 samples of 4567 patients were sampled with replacement. A stepwise logistic regression analysis was applied to every bootstrap sample. If the predictors occurred in more than 50% of the bootstrap models, they were judged to be reliable and were retained in the final model. Unreliable variables, if present, were removed from the final model.

Finally, we internally validated the risk-prediction model by randomly drawing 200 samples each containing 100% of the total number of subjects. The risk-prediction model was applied to each sample to calculate an individual sample area under the ROC curve (AUC) and then the mean and standard error of the mean with 95% confidence intervals (95% CI) for all 200 ROC values.

Two different properties were used to evaluate the predictive accuracy of the model: calibration and discrimination. Calibration was evaluated by the Hosmer–Lemeshow goodness-of-fit method. A statistically non-significant result \((P>0.05)\) suggests that the model predicts accurately on average. In order to get more insight into the model performance across the ranges of patient deciles of risk, we plotted the observed and expected mortality in these risk groups. Discrimination was evaluated by analysis of the AUC. If the area is greater than 0.7, it can be concluded that the model has an acceptable discriminatory power [11] and, consequently, may be used to rank patients into treatment groups to facilitate management.

### 3. Results

#### 3.1. Performance of the external risk scores in our population

The 44 deaths observed, resulted in an overall observed mortality rate of 0.96%. The logistic EuroSCORE predicted a mortality rate of 2.34% \((P<0.001\) vs. observed), the Parsonnet score 4.42% \((P<0.0001\) vs. observed) and the OPR score 1.66% \((P<0.005\) vs. observed). The percent difference between the predicted and observed in-hospital mortality was 58.9% for EuroSCORE \((P<0.001)\), 78.8% for Parsonnet \((P<0.0001)\) and 42.1% for OPR \((P<0.005)\). The exploration of the models at tertiles of risk showed that all three models overestimated mortality significantly in each risk group, except the OPR in the first tertile (Table 2). The Hosmer–Lemeshow test for the three models returned statistically significant results \((P<0.01)\). These results suggest that these risk models do not predict in-hospital mortality accurately and, consequently, their use for risk prediction may not be appropriate in our patient population.

The AUC for the EuroSCORE was 0.754 (95% CI, 0.679–0.828), suggesting that the EuroSCORE may be used in our population only to stratify patients into risk groups. The use of the other two scoring systems is less appropriate, as the AUC for the Parsonnet and the OPR were 0.664 (95% CI, 0.584–0.744) and 0.683 (95% CI, 0.616–0.749), respectively.

#### 3.2. Local risk prediction model for in-hospital mortality

Table 3 summarizes the variables used in the model and their frequency of occurrence (%) in bootstrap analyses, regression coefficients, odds ratio and associated \(P\)-values. Model predictors of in-hospital mortality included: age (increasing), reoperation, peripheral vascular disease, left ventricular dysfunction (EF <40%) and non-elective surgery. All these risk factors occurred in more than 50% of the bootstrap samples, indicating reliability.

The regression model derived significantly predicted the occurrence of in-hospital mortality in this data set \(\chi^2 (5\ d.f.) = 48.45,\ P<0.001\). The correlation between the observed and expected number of deaths was high \((r=0.99)\). The Hosmer–Lemeshow goodness-of-fit test was not statistically significant \((P=0.979)\) and the observed proportion of deaths in each decile risk group tended to conform with the average predicted probability of death in that risk group (Fig. 1). These results indicate that the model accurately predicts in-hospital mortality, both on average and across the ranges of patient deciles of risk and, hence, is suitable for use in all (low to high-risk) patients.
Table 1
Definitions of risk factors in EuroSCORE and local database

<table>
<thead>
<tr>
<th>Risk factor</th>
<th>EuroSCORE definition</th>
<th>Local definition match</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chronic pulmonary disease</td>
<td>Long-term use of bronchodilators or steroids for therapy for the treatment of chronic lung disease</td>
<td>Patient requires pharmacologic therapy for the treatment of chronic pulmonary compromise, or patient has a FEV$_1$ &lt;75% of predicted value</td>
</tr>
<tr>
<td>Extra-cardiac arteriopathy</td>
<td>Any one or more of the following: claudication, carotid occlusion or &gt;50% stenosis, previous or planned intervention on limb arteries or carotids</td>
<td>Patient has peripheral vascular disease as indicated by claudication either with exertion or at rest; amputation for arterial insufficiency; aorta-iliac occlusive disease reconstruction; peripheral vascular bypass surgery, angioplasty or stent; documented AAA, AAA repair, or stent; or non-invasive carotid test with &gt;75% occlusion</td>
</tr>
<tr>
<td>Neurological dysfunction</td>
<td>Severely affecting ambulation or day-to-day functioning</td>
<td>Patient has cerebrovascular disease, documented by any one of the following: unresponsive coma &gt;24 h, CVA, RIND or TIA</td>
</tr>
<tr>
<td>Recent cardiac infarct</td>
<td>&lt;90 days</td>
<td>&lt;30 days</td>
</tr>
</tbody>
</table>

AAA, abdominal aorta aneurism; CVA, cerebrovascular accident; RIND, reversible ischemic neurological deficit; TIA, transient ischemic attack.

The AUC for the model was 0.752 (95% CI, 0.739–0.764), and the mean AUC from the bootstrap re-sampling was 0.752 (95% CI, 0.747–0.758). These results suggest that the risk model has acceptable discriminatory ability, internal validity and stability and, hence, may be used to stratify patients into risk groups for surgical management.

4. Discussion

Currently used risk-score systems have been developed for quite sometime now and do not reflect improved surgical techniques and postoperative patient management advances which occurred in recent times. In addition, they are usually applied without validation to patient populations different from those from which they were derived. We had been aware of the relative incapacity of these risk-score systems to accurately predict in-hospital mortality in our patients subjected to CABG, which was persistently underestimated.

To confirm this assumption, one of the objectives of the present study was to adequately assess the validity of three risk-adjusted predictive models – the EuroSCORE, the Parsonnet score and the OPR score – in predicting in-hospital mortality in our population of coronary surgery patients. To this aim, each of these models’ performance was assessed with regard to discrimination and calibration. We were able to confirm that the three risk-score systems analyzed do not accurately predict outcomes in this group of 4567 patients. They all significantly overestimated total observed outcomes. Additionally, the exploration of risk tertiles showed that all models significantly overestimated mortality at each risk group, except for the OPR in the first tertile. These results suggest the use of these scoring systems for patient advice of risk prediction is not appro-

Table 2
Observed and predicted mortality by risk tertiles

<table>
<thead>
<tr>
<th>EuroSCORE</th>
<th>No. at risk</th>
<th>No. of observed deaths</th>
<th>% observed deaths</th>
<th>No. of predicted deaths</th>
<th>% predicted deaths</th>
<th>% difference</th>
<th>P-value of difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
<td>1448</td>
<td>4</td>
<td>0.28</td>
<td>13.1</td>
<td>0.91</td>
<td>69.5</td>
<td>0.048</td>
</tr>
<tr>
<td>2nd</td>
<td>1597</td>
<td>12</td>
<td>0.75</td>
<td>24.7</td>
<td>1.55</td>
<td>51.4</td>
<td>0.046</td>
</tr>
<tr>
<td>3rd</td>
<td>1522</td>
<td>28</td>
<td>1.84</td>
<td>69.2</td>
<td>4.55</td>
<td>59.5</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Total</td>
<td>4567</td>
<td>44</td>
<td>0.96</td>
<td>107.0</td>
<td>2.34</td>
<td>58.9</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Parsonnet</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1st (0–2)</td>
<td>1230</td>
<td>7</td>
<td>0.57</td>
<td>32.0</td>
<td>2.60</td>
<td>78.1</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>2nd (3–5)</td>
<td>1911</td>
<td>12</td>
<td>0.63</td>
<td>63.8</td>
<td>3.34</td>
<td>81.2</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>3rd (6–28)</td>
<td>1426</td>
<td>25</td>
<td>1.75</td>
<td>135.5</td>
<td>9.50</td>
<td>81.6</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Total</td>
<td>4567</td>
<td>44</td>
<td>0.96</td>
<td>202.3</td>
<td>4.43</td>
<td>78.3</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>OPR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1st (0–1)</td>
<td>2266</td>
<td>9</td>
<td>0.40</td>
<td>9.3</td>
<td>0.41</td>
<td>3.2</td>
<td>1</td>
</tr>
<tr>
<td>2nd (2–3)</td>
<td>1782</td>
<td>25</td>
<td>1.40</td>
<td>41.5</td>
<td>2.33</td>
<td>39.8</td>
<td>0.048</td>
</tr>
<tr>
<td>3rd (4–9)</td>
<td>519</td>
<td>10</td>
<td>1.93</td>
<td>24.8</td>
<td>4.48</td>
<td>59.7</td>
<td>0.015</td>
</tr>
<tr>
<td>Total</td>
<td>4567</td>
<td>44</td>
<td>0.96</td>
<td>75.8</td>
<td>1.66</td>
<td>41.9</td>
<td>0.005</td>
</tr>
</tbody>
</table>
Table 3
Risk model for in-hospital mortality

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficients</th>
<th>P-value</th>
<th>Bootstrap frequency, %</th>
<th>Odds ratio</th>
<th>95% CI (OR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (per one year increase)</td>
<td>0.052</td>
<td>0.006</td>
<td>97.8</td>
<td>1.054</td>
<td>1.015 - 1.094</td>
</tr>
<tr>
<td>Peripheral vascular disease</td>
<td>1.575</td>
<td>&lt;0.001</td>
<td>71.2</td>
<td>4.831</td>
<td>2.590 - 9.010</td>
</tr>
<tr>
<td>LV dysfunction (EF&lt;40%)</td>
<td>0.688</td>
<td>0.055</td>
<td>67.3</td>
<td>1.989</td>
<td>0.985 - 4.016</td>
</tr>
<tr>
<td>Non-elective surgery</td>
<td>1.201</td>
<td>0.002</td>
<td>61.3</td>
<td>3.323</td>
<td>1.556 - 7.097</td>
</tr>
<tr>
<td>Reoperation</td>
<td>1.617</td>
<td>0.009</td>
<td>53.5</td>
<td>5.040</td>
<td>1.486 - 17.090</td>
</tr>
<tr>
<td>Intercept</td>
<td>-8.601</td>
<td>&lt;0.001</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Model: χ² [5 d.f.] = 48.45, P<0.001.

appropriate in our population. However, the discriminatory ability of the EuroSCORE was good, with an AUC of 0.754, suggesting that this risk-score may be used in our population to stratify patients into risk groups for treatment management.

Consequent to these findings, confirming our previous assumptions, the main goal of this study was the development of our own risk model for our patient population undergoing CABG surgery, which could be used as an instrument to provide information to clinicians and patients about the risk of surgical mortality.

The risk factors included in our risk model were: age, reoperation, peripheral vascular disease, left ventricular dysfunction and non-elective surgery. The main risk factors observed here remain consistent with the findings in most previously published risk models for CABG mortality [10]. On the other hand, and in contrast to what has been found in other studies [12–14], population variables such as female sex, renal dysfunction and diabetes mellitus, did not emerge as independent risk factors in this study. The prediction model demonstrated acceptable discriminatory ability and accurately predicts in-hospital mortality, both on average and across the ranges of patient deciles of risk.

The end-point of the study was in-hospital mortality. Although it represents one of the most widely reported metrics to assess death after CABG, it may be a too short interval for the evaluation of early risk. Nevertheless, and in the context of the present study, we believe that the more important issue, other than the specific measure used, is the ability to measure and validate it conveniently and accurately. The mortality risk predicted by the EuroSCORE was only 2.34%. This result places this patient cohort in a low risk profile, which means that any inference must be reduced to the center where it was developed, possibly limiting the applicability to others.

Although there is no consensus on sample size, as a rule of thumb in studies deriving multivariable prognostic models, ten or more events per variable are usually required in order to get a robust estimation of the coefficients. The ratio of events to risk factors included in our local model was approximately 9–1 (44 events; 5 variables), therefore, the data of the multivariate analysis should be interpreted with caution.

In our database, some of the variables selected for analysis (ejection fraction, hematocrit, cardiothoracic ratio) were codified as categorical instead of continuous variables and, consequently, this fact constitutes one limitation to the process of correct model building.

5. Conclusion

We believe that scores described in published studies are often not correct because the scoring system had not been appropriately validated for the respective populations. We developed a risk-prediction model that can be used as an instrument to provide information to clinicians and patients about the risk of in-hospital mortality in our patient population awaiting CABG surgery. Naturally, it is for our own use and is not intended for use in other patient populations.

Appendix A

Risk factors included in the data-base used for the calculations

Age, gender, body mass index (BMI), diabetes (no/yes; history of diabetes treated with oral agents or insulin), hypertension (no/yes; blood pressure exceeding 140/90 mmHg, or a history of high blood pressure, or the need of antihypertensive medications), renal failure (none or functioning transplant/creatinine > 2.0 mg/dl and no dialysis dependency), recent smoking (no/yes; less than four weeks of surgery), anemia (no/hematocrit ≤ 34%), cardio-megaly (no/cardiothoracic ratio > 0.50 on a chest X-ray-film), chronic pulmonary disease (no/yes), peripheral vascular disease (no/yes), cerebrovascular disease (no/yes), recent myocardial infarction (no/yes), unstable angina (no/yes), angina CCS class III or IV (no/yes), left main...
disease (no/yes), three vessel disease (no/yes), reoperation (no/yes), left ventricular dysfunction (no/ejection fraction <40%), non-elective surgery (no/patient requires urgent or emergent surgery), intra-aortic balloon pump (no/preoperative intra-aortic balloon pump for hemodynamic reasons).

References


Conference discussion

Dr. P. Kolh (Liege, Belgium): That is a very nice study. Recently a colleague from Sweden published a very comprehensive study in the European Heart Journal (Nilsson J. et al. 2006;27:867–874) comparing 19 models that would predict risk in cardiac surgery, and the best one for coronary artery surgery was the EuroSCORE, followed by the Cleveland Clinic and the New York State scores. The Parsonnet score did not perform well, and it could be expected because the Parsonnet score is relatively old now, about 20 years old, I have two comments and questions.

The first one is a question. Do you plan to validate your model using another population? Because now you have developed a model and you have validated your model with your own population, but you would need to see whether it works in other population settings.

The second one is, if I remember correctly, in one of your slides you are using in-hospital mortality while most models use 30-day mortality. It is an important difference, because if you decrease the length of hospital stay, for a total number of patients who died within 30-day, more patients would die in the interval between hospital discharge and the 30-day endpoint. So I would appreciate if you could comment on that also.

Dr. Antunes: I’ll start with that question first. We just wanted this study as an exercise of assessing our own performance and because most of our patients go out to their cardiologists, it was difficult for this analysis to try and get all the follow-up data on the 4,500 patients. That is why we used that. We recognize that it will underestimate the mortality, but the curves are pretty parallel. And if you see our initial comparison, it also shows that the EuroSCORE, although being the best performer, was a little bit away from our own observed and expected rates of mortality.

To answer your first question, as we developed the model we observed that as our experience progressed, the curves started to diverge again. So we need to recalibrate these models and time and again. And that is the problem with the currently used models is that they were established some 10, 15 or 20 years ago and they were not recalibrated for current needs.

Dr. P. Kappetein (Rotterdam, The Netherlands): Great presentation and I fully agree with the previous discussant that the models that are currently available are not so valid anymore. I wonder if everybody shouldn’t use a model they develop in their own institution. There are now many papers in literature that show that EuroSCORE gives a higher predicted than observed mortality and many authors present their own scoring system. So my question is, do you think that we now should use the Coimbra score instead of the EuroSCORE or that we should develop a score for our own institution?

Dr. Antunes: No, the message I want to bring is that these commonly accepted risk scores do not always predict accurately your own internal results, specific of your local institution. This model we developed is for internal use only, so that we keep a record of our own performance. We do not intend to suggest to anybody to use the score, because obviously the populations are different and the methodologies are different. We cannot compare our own performance, and we need to keep a track on that, if we constantly use a model that shifts far away from our observed circumstances. That is all.

Dr. K. Hekmat (Ulm, Germany): Congratulations on your score because there are only five variables and I think that is very nice for all the residents who have to do this scoring. I have just problems with two of the variables. One is peripheral vascular disease. You didn’t define it, because you can have a different extent. And the other one is also the ejection fraction, because you don’t have it on all the patients, and the same is also true for the peripheral vascular disease. So if you don’t have data on these two variables, I think you can get problems with the score.

Dr. Antunes: No, we do have that data, and the presentation is limited to five minutes. The paper, if it is published, and I hope so, will have the definitions of all those 22 variables that we have here. But, just for your own information, LV dysfunction was defined as less than 40% ejection fraction, and peripheral vascular disease was diffuse disease in more than one territory.

Dr. Hekmat: And you have data on all the patients, 100%?

Dr. Antunes: You can’t have complete data on 4,500 patients, but I would say more than 95%, because this is a prospectively collected database. I can’t guarantee that all the surgeons have put in all the variables, but pretty close to that.