Anaesthesia
Journal of the Association of Anaesthetists of Great Britain and Ireland

Anaesthesia, 2008, **63**, pages 705–713 doi:10.1111/j.1365-2044.2008.05478.x

Predicting adverse outcomes of cardiac surgery with the application of artificial neural networks

S-Y Peng[1,*] and S-K Peng[2]

1 PhD student, Institute of Biomedical Informatics, National Yang-Ming University, Taipei, Taiwan
2 Attending, Department of Anesthesiology, Taichung Veterans General Hospital, Taichung, Taiwan

Summary

Risk-stratification models based on pre-operative patient and disease characteristics are useful for providing individual patients with an insight into the potential risk of complications and mortality, for aiding the clinical decision for surgery vs non-surgical therapy, and for comparing the quality of care between different surgeons or hospitals. Our study aimed to apply artificial neural networks (ANN) models to predict mortality and morbidity after cardiac surgery, and also to compare the efficacy of this model to that of the logistic regression model and Parsonnet score. The accuracy of the ANN, logistic regression and Parsonnet score in predicting mortality was 83.8%, 87.9% and 78.4%. The accuracy of the ANN, logistic regression and Parsonnet score in predicting major morbidity was 79.0%, 74.3% and 68.6%. The area under the receiver operating characteristic curves (AUC) of the ANN, logistic regression and Parsonnet score in predicting in-hospital mortality were 0.873, 0.852 and 0.829. The AUCs of the ANN, logistic regression and Parsonnet score in predicting major morbidity were 0.852, 0.789 and 0.727. The results showed the ANN models have the best discriminating power in predicting in-hospital mortality and morbidity among these models.

Correspondence to: Shih-Kuei Peng
**Present address: No.108-3, Sec. 2, Fuxing Rd., South District, Taichung City 402, Taiwan (R.O.C.).*
E-mail: peng.kuei@msa.hinet.net
Accepted: 6 January 2008

Advances in medical treatment, the advent of thrombolytic therapy, and the availability of a range of percutaneous angiographic interventions has changed the profile of patients referred for cardiac surgery [1, 2]. Referred patients tend to be older, with a substantial increase in the proportion of high-risk patients [3–5]. Risk-stratification models based on pre-operative patient and disease characteristics are useful for providing patients with an insight into the potential risk of complications and mortality, for aiding the clinical decision for surgery vs non-surgical therapy, and for comparing the quality of care between different surgeons and hospitals.

Parsonnet and colleagues first designed a risk stratification model for evaluation of the mortality after cardiac surgery in 1989 [6]. The Parsonnet score is a simple scoring system which derived from large international statistics tries, on the basis of pre-operative risk factors, to assess and predict the mortality of patients with coronary and heart valve operations. Thereafter, numerous multifactorial risk scores have been developed to predict the outcomes following cardiac surgery [7–12]. Most of the scoring systems have been designed to predict mortality, but postoperative morbidity is recognised as a major determinant of hospital costs and quality of life after surgery [13]. Relatively few studies have involved the prediction of postoperative morbidity following cardiac surgery [8, 14].

Logistic regression models the probability of some event occurring as a linear function of a set of predictor variables. The actual state of the dependent variable is determined by looking at the estimated probability.

The artificial neural network (ANN) is a computational model with parallel non-linear processing elements arranged in highly interconnected networks, emulating complex human thought processes such as adaptive learning, optimisation and reasoning. Clinical studies

have suggested the superiority of ANN, as compared to other statistical models, for the classification or prediction of morbidity [15, 16]. The ability of the ANN to process more information in the context of multidimensionality of complex data to a large extent explains the superiority of ANN as a predicting model. Our study aimed to apply a separate ANN model to predict each of the morbidity and mortality following cardiac surgery. We also compared the discrimination of our ANN models to those of the logistic regression model and Parsonnet score in prediction of in-hospital mortality and morbidity.

Methods

We retrospectively studied 952 anonymised case records of adult patients who underwent cardiac surgery during a three year period (from August 2004 to December 2006) at the Taichung Veterans General Hospital, Taiwan. No formal ethical permission was deemed necessary for this exercise. Patients who underwent cardiac transplantation or surgery for congenital heart disease were excluded due to the small numbers of cases. Patient records were reviewed by two anesthesiologists and the pre-operative patient characteristics were noted. The data collected included age, gender, body height, body weight, left ventricular ejection fraction (LVEF), renal function (divided into three groups: normal; dialysis dependency; and renal insufficiency, i.e. serum creatinine level exceeding 2.0 $mg.dl^{-1}$ without dialysis), priority of surgery (emergency, defined as surgery within 24 h of cardiac catheterisation; and urgency, defined as diagnosis and surgery within the same admission), surgical procedure (simple, defined as a single procedure; complex, defined as combined valve and coronary artery surgery, multiple valve surgery or surgery for ascending aortic aneurysm), prior cardiac surgery, dyspnoea class (as per New York Heart Association, NYHA), prior myocardial infarction (MI) within the previous 6 weeks, intra-aortic balloon pump (IABP) before operation, cerebrovascular disease, hypertension (defined as systolic blood pressure > 160 mmHg, or diastolic > 90 mmHg; or receiving active treatment for hypertension), ventricular aneurysm, pulmonary hypertension (defined as systolic pulmonary artery pressure above 60 mmHg), aortoventricular pressure gradient, chronic obstructive pulmonary disease (COPD), diabetes mellitus (DM), catastrophic states (e.g., acute structural defect, cardiogenic shock and acute renal failure) and other rare circumstances (e.g., paraplegia, pacemaker dependency and severe asthma).

The measure we used to select the predictors to build logistic regression and ANN models followed the recommendations of Dreiseitl et al. [17]. For logistic regression models, it is possible to test the statistical significance of the coefficients in the model; these tests can be used to build models incrementally. The three most common approaches are to start with an empty model and successively add covariates (forward selection), to start with the full model and remove covariates (backward selection), or to both add and remove variables (stepwise selection). Due to the non-linear nature of ANN, the conventional statistical tests used to test significance of predictors that are used in logistic regression cannot be applied here. Instead, one can use automatic relevance determination or sensitivity analysis to assess the relative importance of predictors. The final results thus do depend upon the chosen measure we used to select predictors.

Outcomes

The main outcomes of our study included in-hospital mortality and major postoperative morbidity. In-hospital mortality was defined as a death occurring during the hospital stay. Major morbidity was defined using criteria previously reported by Dupuis et al. [14], and described as follows:

1 Cardiovascular: low cardiac output, hypotension, or both, treated with IABP, with two or more intravenous inotropes or vasopressors for more than 24 h, or with both; malignant arrhythmia (asystole and ventricular tachycardia or fibrillation) requiring cardiopulmonary resuscitation, anti-arrhythmia therapy, or automatic cardiodefibrillator implantation.
2 Respiratory: mechanical ventilation for more than 48 h, tracheostomy, reintubation.
3 Neurological: focal brain injury with permanent functional deficits, irreversible encephalopathy.
4 Renal: acute renal failure requiring dialysis.
5 Infection: septic shock with positive blood cultures, deep sterna or wound infection requiring intravenous antibiotics, surgical debridement, or both.
6 Other: any surgery or invasive procedure necessary to treat a postoperative adverse event (i.e. graft failure or re-bleed) associated with the initial cardiac surgery.

Risk stratification of patients

Patients were randomly allocated to a training set and a validation set. The training set included two-thirds of all patients who had undergone cardiac surgery (n = 637). The remainder of the study population formed the validation set (n = 315). The multifactorial risk score was determined for each patient, in accordance with the risk indexes developed for general cardiac surgical populations by Parsonnet et al. (Table 1). The Parsonnet score was calculated for the training set, separately for the outcomes of mortality and morbidity.

Table 1 The Parsonnet score for the prediction of outcomes after cardiac surgery.

Parsonnet score	
Age (years)	
70–74	7
75–79	12
> 80	20
Emergency after cath.	10
LV function	
EF 39–49%	2
EF < 30%	4
Surgical characteristics	
MVR or AVR	5
CABG + valve	2
Reoperation	
1st	5
2nd	10
Female gender	1
Dialysis dependency	10
Systolic PAP > 60 mmHg	8
Diabetes	3
Morbid obesity	3
A-V gradient > 120 mmHg	7
Pre-operative IABP	2
Ventricular aneurysm	5
Hypertension	3
Catastrophic states	*10–50
Rare circumstance	*2–10
Maximum score	158

*Points are allocated subjectively by the observer.
Cath., coronary angiogram; LV, left ventricle; EF, ejection fraction; MVR, mitral valve replacement or repair; AVR, aortic valve replacement; CABG, coronary artery bypass grafting; PAP, pulmonary artery pressure; A-V, aortoventricular; IABP, intra-aortic balloon pump.

Logistic regression evaluates the probability of an event occurring as a linear function of a set of predictor variables. The actual state of the dependent variable is determined by the estimated probability. Pre-operative patient characteristics were used to build logistic regression models. The predictors were initially chosen by univariate analysis ($p < 0.2$). The picked predictors were then screened by backwards stepwise selection ($p < 0.05$ to retain in the logistic regression model). We used the training set to develop separate logistic regression models for each of the two outcomes. The validation set was subjected to the Parsonnet score and logistic regression.

Construction of ANN models

Neural computation was performed on an IBM compatible Pentium 4 computer running at 2.13 GHz. STATISTICA 7.0 (StatSoft, Inc., Tulsa, OK, USA) was used to build ANN models with the training set. Three-quarters of the training set ($n = 478$) were optimised to fit the ANN model, and the remaining patients ($n = 159$) were used as a selection subset to estimate the prediction error for stop training to mitigate over-learning and over-fitting. The predictors were entered into the ANN as dichotomous 'yes/no' input variables or numerical variables. Mortality or morbidity was entered as a dichotomous 'yes/no' output variable. A multilayer perception (MLP) ANN with a back propagation algorithm was used to train the predictive model. The MLP network consists of layers of virtual neurons in which neurons from contiguous layers are linked to each other by weighted connections. Each neuron receives an input of a weighted sum of the output of all neurons in the previous layer. It then processes this input using a non-linear transfer function and the processed output forms the input to neurons in the next layer [17, 18].

During network training, a prediction was made and was correlated with the observed outcome. If the ANN predicted the outcome incorrectly, the error difference was calculated and the error was back-propagated through the network to adjust the connection weights to more closely match input and output data. Input variables were pruned if the fan-out weights of the variable-associated input layer neurons were below the threshold of 0.05. The weightings were self-selected by the process. The training process was repeated 5000 times, and the model with the lowest prediction error in the selection subset was retained. The configuration of the ANN was chosen by the software (the retained ANN had lowest number of neurons to minimise predictive error). The selected ANN consisted of one input layer, one hidden layer and one output layer (Fig. 1). The ANN model used to predict mortality had 29 neurons in the input layer, 15 neurons in hidden layer and one neuron in the output layer. The ANN model for morbidity had 31 neurons in the input layer, 15 neurons in the hidden layer and one neuron in the output layer. Sensitivity analysis was used to evaluate the importance of predictors. The predictive error ratio, i.e. the relation between predictive errors for a given predictor to the total network errors for all predictors, was calculated for each input predictor.

On completion of the training process, the networks were tested with the validation set, which had not been involved in the training, and whose mortality and morbidity outcomes were not known to the ANN. It is important to emphasise that even with this process, we could not know if the ANN model was fully optimised. Because the ANN can generate infinite sets of models, we can only choose the best model from the training course. The bigger the training set we used and the more training times the ANN underwent, the more likely we were to get a better model.

Statistical analysis

Common measures of discrimination are sensitivity, specificity, accuracy and the area under the receiver operating characteristic (ROC) curve. ROC analysis

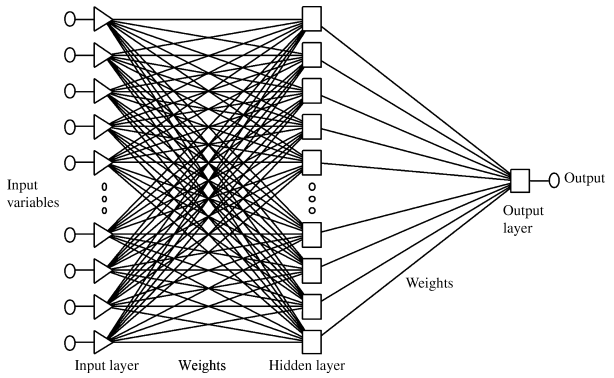

Figure 1 Illustration of the artificial neural network (ANN) model. Input variables: predictors (e.g. age, gender, etc.). Input layer: traing set, testing set. Hidden layer: present predictor's internal relationship. Output layer: normalise output, competitive output, competitive learning. Weights: developed by the ANN to connect between layers. Output: outcome (mortality or morbidity).

estimates a curve that describes the inherent trade-off between sensitivity and specificity of a prediction tool. The area under the ROC curve (AUC) is a particularly important metric for evaluating prediction tools because it indicates the average sensitivity over all possible specificities. The AUC may range from zero to one, with an area of 1.0 representing perfect discrimination, and an area of 0.5 representing what can be expected by chance alone. An AUC between 0.7 and 0.8 was classified as 'acceptable', and between 0.8 and 0.9 as 'excellent'. A cut-off value, corresponding with the highest accuracy (minimal false negative and false positive results), was selected to report the sensitivity and specificity of each prediction tool. The performance of these models was compared with each other by using pairwise comparison of ROC curves. The statistical significance for pairwise comparisons among three models was defined as p value < 0.017 by the Bonferroni correction.

Calibration with goodness of fit can evaluate the degree of correspondence between outcome frequencies produced by prediction models and actual observation. This was measured by the Hosmer–Lemeshow goodness-of-fit statistic, which divides subjects into deciles based on predicted probabilities and then computes a chi-square statistic from observed and expected frequencies. A statistically good fit was defined at the 5% probability level.

Results

Nine hundred and fifty-two patients undergoing cardiac surgery during a 3-year period were retrospectively studied. The characteristics of these patients are presented in Table 2. The in-hospital mortality and major morbidity rates in the training set ($n = 637$) were 10.2% and 46.6%, respectively. Comparable patient characteristics, mortality (11.7%), and morbidity (48.6%) rates were identified for the validation group ($n = 315$). The overall in-hospital mortality and morbidity of our study population were 10.7% and 47.3% respectively.

Table 2 Patient characteristics.

	Training set ($n = 637$)	Validation set ($n = 315$)
Age (years)	63.2 (13.6)	64.8 (13.8)
Height (cm)	163.1 (8.0)	161.9 (8.5)
Body weight (kg)	65.4 (11.3)	64.7 (11.5)
Female	154 (24%)	94 (30%)
LV function (ejection fraction)		
< 20%	23 (4%)	11 (4%)
20–34%	108 (17%)	63 (20%)
35–50%	208 (33%)	113 (36%)
> 50%	298 (47%)	128 (41%)
Renal function		
Dialysis dependency	23 (4%)	13 (4%)
Renal insufficiency	65 (10%)	36 (11%)
Normal	549 (86%)	266 (84%)
Surgical priority		
Emergency	96 (15%)	42 (13%)
Urgency	69 (11%)	26 (8%)
Elective	472 (74%)	247 (78%)
Surgical procedure		
Complex	132 (21%)	82 (26%)
Simple	505 (79%)	233 (74%)
Reoperation	28 (4%)	12 (4%)
Cardiogenic shock	39 (6%)	17 (5%)
Congestive heart failure		
NYHA class 1	202 (32%)	92 (29%)
NYHA class 2	277 (44%)	135 (43%)
NYHA class 3	105 (17%)	56 (18%)
NYHA class 4	53 (8%)	32 (10%)
MI within 6 weeks	65 (10%)	35 (11%)
Pre-op IABP support	40 (6%)	14 (4%)
Cerebrovascular disease	69 (11%)	33 (11%)
Hypertension	405 (64%)	196 (62%)
LV aneurysm	28 (4%)	16 (5%)
Pulmonary hypertension	215 (34%)	108 (34%)
A-V pressure gradient ≥ 120 mmHg	3 (1%)	2 (1%)
COPD	39 (6%)	12 (4%)
Diabetes	221 (35%)	104 (33%)
Catastrophic states	88 (14%)	42 (13%)
Rare circumstances	33 (5%)	22 (7%)
In-hospital mortality	65 (10%)	37 (12%)
Major morbidity	297 (47%)	153 (49%)

Data are mean (SD) or numbers (%).
LV, left ventricle; NYHA, New York Heart Association; A-V, aortoventricular; LV, left ventricle; IABP, intra-aortic balloon pump; COPD, Chronic obstructive pulmonary disease.

Table 3 shows the predictors used in the Parsonnet score, logistic regression and ANN. Significant risk factors selected to build the logistic regression model for predicting mortality included age, LV function, renal function, surgical priority, re-operation, cardiogenic shock, dyspnoea class, MI in the previous 6 weeks,

Table 3 Predictors used by Parsonnet score, logistic regression and artificial neural network (ANN).

	In-hospital mortality			Morbidity		
	Parsonnet score	Logistic regression	ANN	Parsonnet score	Logistic regression	ANN
Age	○	○	○	○	○	○
Height						
Body weight						
Gender	○		○	○		○
LV function	○	○	○	○	○	○
Renal function	○	○	○	○	○	○
Surgical priority	○	○	○	○	○	○
Surgical procedure	○		○	○	○	○
Reoperation	○	○		○		○
Cardiogenic shock		○	○		○	
Congestive heart failure		○	○		○	○
MI within 6 weeks		○	○		○	○
Pre-op IABP support	○	○	○	○	○	
Cerebral vascular disease		○	○		○	○
Hypertension	○	○	○	○	○	○
LV aneurysm	○	○	○	○	○	○
Pulmonary hypertension	○		○	○	○	○
A-V pressure gradient ≥ 120 mmHg	○			○		
COPD					○	○
Diabetes	○	○	○	○	○	○
Catastrophic states	○	○	○	○	○	○
Rare circumstances	○	○		○	○	○

MI, myocardial infarction; LV, left ventricle; A-V, aortoventricular; IABP, intra-aortic balloon pump; COPD, chronic obstructive pulmonary disease.

pre-operative IABP use, cerebrovascular disease, hypertension, LV aneurysm, diabetes, catastrophic states and rare circumstances. Risk factors for the logistic regression model to predict morbidity included age, LV function, renal function, surgical priority, complexity of surgery, cardiogenic shock, dyspnoea class, MI in the previous 6 weeks, pre-operative IABP use, cerebrovascular disease, hypertension, LV aneurysm, pulmonary hypertension, COPD, diabetes, catastrophic states and rare circumstances.

The predictors chosen by the ANN model to predict mortality included age, gender, LV function, renal function, surgical priority, complexity of surgery, cardiogenic shock, dyspnoea class, MI in the previous 6 weeks, pre-operative IABP use, cerebrovascular disease, hypertension, LV aneurysm, pulmonary hypertension, diabetes and catastrophic states. The predictors retained in the model for predicting morbidity included age, gender, LV function, renal function, surgical priority, complexity of surgery, re-operation, dyspnoea class, MI in the previous 6 weeks, cerebrovascular disease, hypertension, LV aneurysm, pulmonary hypertension, COPD, diabetes, catastrophic states and rare circumstances. The sensitivity analysis of the predictors is presented in Table 4. The contributions of different input variables to predicting the outcomes are ranked in order of descending importance.

The ANN had an overall accuracy of 83.8% for predicting in-hospital mortality. The sensitivity was 81.1%, and specificity was 84.2%. The accuracy of logistic regression was 87.9%; the sensitivity was 67.6%, with a specificity of 90.7%. The accuracy of the Parsonnet score was 78.4%; the sensitivity was 78.4%, with a specificity of 78.4%. The accuracy of the ANN, logistic regression and Parsonnet score in predicting major morbidity was 79.0%, 74.3% and 68.6%; the sensitivity was 71.9%, 67.3%, 53.6%; and the specificity was 85.8%, 80.9% and 82.7% respectively. Table 5 shows the ability of the ANN, logistic regression and Parsonnet score to predict the major outcomes in the validation set.

The AUCs of the ANN, logistic regression and Parsonnet score in predicting in-hospital mortality were 0.873, 0.852 and 0.829 (Fig. 2). The AUCs of the ANN, logistic regression and Parsonnet score in predicting major morbidity were 0.852, 0.789 and 0.727 (Fig. 3). The discriminating powers of these models were compared using pairwise analysis of ROC curves (Table 6). We found that there were no statistically significant differences between the ANN, logistic regression and Parsonnet score in predicting in-hospital mortality. However, the ANN model performed significantly better than the other models in predicting morbidity. All of these models had a statistically good fit (at 5% probability level) related to the actual outcomes.

Table 4 Sensitivity analysis of predictors in the artificial neural network models.

	In-hospital mortality		Morbidity	
	Predictive error ratio	Rank	Predictive error ratio	Rank
Age	1.0176	7	1.0000	11
Height	N/A	N/A	N/A	N/A
Body weight	N/A	N/A	N/A	N/A
Gender	0.9916	16	1.0808	1
LV function	1.0876	1	1.0606	2
Renal function	1.0700	3	0.9964	12
Surgical priority	1.0406	4	0.9910	14
Surgical procedure	1.0160	8	0.9863	15
Reoperation	N/A	N/A	1.0140	7
Cardiogenic shock	1.0110	11	N/A	N/A
Congestive heart failure	1.0749	2	1.0319	3
MI within 6 weeks	1.0006	14	1.0258	6
Pre-op IABP support	1.0026	13	N/A	N/A
Cerebral vascular disease	1.0379	5	0.9794	17
Hypertension	1.0140	9	1.0073	9
LV aneurysm	1.0089	12	0.9942	13
Pulmonary hypertension	0.9937	15	1.0024	10
A-V pressure gradient ≥ 120 mmHg	N/A	N/A	N/A	N/A
COPD	N/A	N/A	1.0318	4
Diabetes	1.0181	6	0.9837	16
Catastrophic states	1.0122	10	1.0298	5
Rare circumstances	N/A	N/A	1.0113	8

The predictive error ratio was calculated for each input variable according to their degree of validity. The ratio was expressed as the relation between the predictive errors for the model with a removed given input variable and the total network errors calculated based on all input variables. The rank represented that the contributions made by different input variables in predicting the outcome were ranked in order of descending importance.
LV, left ventricle; MI, myocardial infarction; A-V, aortoventricular; IABP, intra-aortic balloon pump; COPD, Chronic obstructive pulmonary disease; N/A, Not available, since the predictor has been pruned by the artificial neural networks.

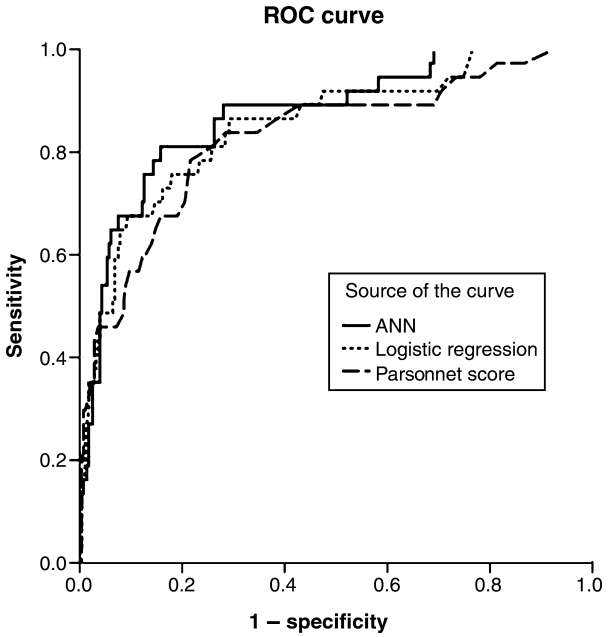

Figure 2 Receiver operating characteristic (ROC) curves of the artificial neural network, logistic regression, and Parsonnet score for predicting in-hospital mortality.

Discussion

Risk stratification for cardiac surgery has gained increasing importance in recent years. Quantification of operative risk helps to: (i) weigh the risk of surgical vs conservative treatment; (ii) inform patients of their peri-operative risk; (iii) allow comparison of outcomes and cost analyses between institutions and surgeons; (iv) facilitate quality monitoring by allowing comparison of outcomes from year to year, or before and after a change in practice; and (v) aid in clinical decision-making based on risk and predicted cost [19]. However, the risk

Table 5 Comparison of predictive performance of artificial neural network, logistic regression and Parsonnet score.

	Accuracy	Sensitivity	Specificity	Area undue ROC curve (AUC)
In-hospital mortality				
ANN	83.8	81.1 [64.8–92.0]	84.2 [79.3–88.3]	0.873 [0.831–0.908]
Logistic regression	87.9	67.6 [50.2–82.0]	90.7 [86.6–93.8]	0.852 [0.808–0.889]
Parsonnet score	78.4	78.4 [61.8–90.1]	78.4 [73.1–83.1]	0.829 [0.783–0.869]
Major morbidity				
ANN	79.0	71.9 [64.1–78.9]	85.8 [79.5–90.8]	0.852 [0.808–0.889]
Logistic regression	74.3	67.3 [59.3–74.7]	80.9 [74.0–86.6]	0.789 [0.739–0.833]
Parsonnet score	68.6	53.6 [45.4–61.7]	82.7 [76.0–88.2]	0.727 [0.674–0.775]

The accuracy is the degree of conformity of a predicted outcome to its actual status. The sensitivity is the proportion of true positives of all cases with adverse outcome in the population. The specificity is the proportion of true negatives of all cases without adverse outcome in the population. ROC, receiver operating curve; AUC, area under the ROC curve; LR+, positive likelihood ratio; LR–, negative likelihood ratio; ANN, artificial neural network.

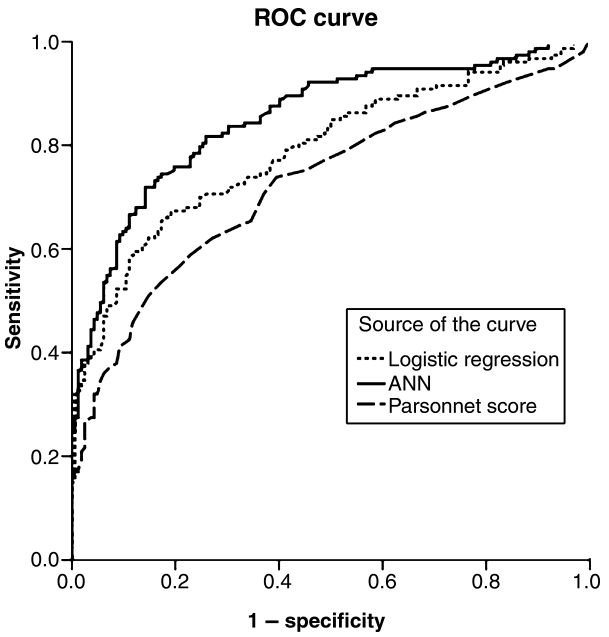

Figure 3 Receiver operating characteristic (ROC) curves of the artificial neural network, logistic regression, and Parsonnet score for predicting major morbidity.

stratification systems most commonly used by anesthesiologists and surgeons, such as the Goldman index, the Acute Physiologic and Chronic Health Evaluation (APACHE) score and the American Society of Anesthesiologists (ASA) health status classification, were not specially designed for cardiac surgery. Therefore, an accurate and objective evaluation tool for assessing peri-operative risk for cardiac surgery has important consequences for patients, physicians and administrators. Although mortality has been referred to as the most important performance indicator in cardiac surgery, and is the most frequently reported outcome parameter in evaluating risk scores, major morbidity can profoundly influence both health care cost and quality of life. Prolonged duration of intensive care unit (ICU) and total postoperative stay resulted from major complications would substantially increase the cost of surgery. In addition, the number of available ICU beds may be limited by a large number of long-stay patients or, more commonly, because of a shortage of trained cardiac ICU nurses. It is necessary to identify a group of patients with a high probability of requiring < 24 h ICU stay in order that some operations can be carried out every day. Prediction of major morbidity is beneficial in prospective planning of resource utilisation. Although we demonstrated a higher AUC with ANN, and we believe that this could be translated into reduced costs of surgery and hospital stay, we cannot precisely confirm this. There is of course no evidence that a unit gain in AUC is readily translatable into a known cost or morbidity saving.

Most risk stratification systems were developed using traditional biostatistical methods [6, 8, 9, 11, 13, 20]. The evolution of ANN has become more satisfactory in medical research than before. A speculative benefit of this artificial intelligence is its power to identify complex correlative interactions among clinical data and final diagnosis. The ANN is a dynamic model; with each new patient, the model back-propagates and checks data with an error-minimisation function, re-adjusting hidden weights to improve predictive accuracy. Thus, as data for more patients were entered into the model, self-learning and error correction by back-propagation, and in turn, predictive accuracy, improved progressively. The major advantage of using an ANN for predicting outcomes following cardiac surgery is that neural networks train themselves without much human intervention. The ANN model is an algorithm that can be used to perform non-linear statistical modeling and provides a new alternative to logistic regression, the most commonly used method for developing predictive models for dichotomous outcomes in the field of medicine. The advantages of the ANN include the need for less formal statistical training, ability to identify complex non-linear relationships between dependent and independent variables,

Table 6 Pairwise comparison of ROC curves.

Pairwise comparison	Difference between AUCs	Standard error	95% CI	p-value
In-hospital mortality				
ANN vs logistic regression	0.021	0.027	−0.033–0.075	0.444
ANN vs Parsonnet score	0.044	0.040	−0.035–0.122	0.275
Logistic regression vs Parsonnet score	0.023	0.035	−0.046–0.091	0.515
Major morbidity				
ANN vs logistic regression	0.063	0.026	0.013–0.114	< 0.017
ANN vs Parsonnet score	0.125	0.030	0.066–0.184	< 0.017
Logistic regression vs Parsonnet score	0.062	0.023	0.017–0.108	< 0.017

ANN, artificial neural network; ROC, receiver operating curve; AUC, area under the ROC curve; CI, confidence interval.

ability to detect all possible interactions between predictor variables, and availability of multiple training algorithms. Disadvantages include its 'black box approach', greater computational burden, proneness to over-fitting and the empirical nature of model development.

Only a few studies using ANNs to predict mortality after cardiac surgery have been published [21–25]. Most of these studies were based on coronary artery bypass graft (CABG)-only patients [22, 23, 25]. Our study underlines the applicability of ANN models for assessing in-hospital mortality and major morbidity after cardiac surgery. We selected the Parsonnet score for validation because the variables used by the model were routinely and completely recorded in our database. In the previous studies of Orr [24], Tu [25], and Lippmann [23], the ANNs showed a performance equivalent to that of logistic regression and other prediction models. Our study showed similar results in predicting in-hospital mortality; however, the performance of the ANN seemed to be better than other models for predicting morbidity. It is unclear why ANN might perform better for morbidity but not for mortality than the other methods. One reason is that this is a chance result. Another possibility relates to spurious accuracy since we forced a classification of morbidity into a dichotomous scale when in reality it is a continuous scale. It is possible that, in a situation where mortality is relatively high, such as cardiac surgery, it is easier even for a sub-optimal prediction method (e.g. logisitic regression or Parsonnet) to predict this common event reasonably accurately; on the other hand these perform less well for morbidity. A final possibility is that morbidity is inherently more 'complex' a phenomenon than is mortality (e.g. the latter is binary and there are no diagnostic dilemmas). Since ANNs are computer models composed of parallel, non-linear computational neurons arranged in highly interconnected layers, they can define relationships among input data that are not apparent when using traditional statistical approaches, and they can use these relationships to improve accuracy. Hence, neural nets have substantial power to identify patterns in complex datasets. Due to the non-linear nature of ANNs, the statistical tests for variable significance that are used in logistic regression cannot be applied here [17]. Therefore, we used sensitivity analysis to assess the importance of input predictors for the classification result. However, we cannot know why some variables might be rejected for mortality but not for morbidity or vice versa (Table 3) since one of the drawbacks of ANN is that it is considered as 'black box' in its establishment of the model, hiding any information on the relationship between the input and the output.

Our study used models which limited the predicting variables to those known before the operation occurred. In fact, some factors such as surgical technique, quality of care, echocardiographic and catheterisation data related to the severity of cardiac disease, extracorporal circulation time, therapeutic interventions and chance happenings, not related to pre-operative patient characteristics (e.g. surgical error or medical error), may also influence the outcomes of cardiac surgery [26, 27]. It is impossible to include all of these variables in a model for pre-operative risk stratification, although their inclusion may improve predictions.

There were some limitations in our study. Firstly, the accuracy of models can often be assessed by introducing 'confounding factors'. Unfortunately this was not possible here as this would have inevitably led to poorer accuracy of the ANN [17]. Thus a limitation of the ANN is that, when it is inaccurate, it is not possible to know if this is because of its own analytic process, or because of the factors that were fed into the ANN. Secondly, we defined the outcome of morbidity as yes/no and it might explain a 'spurious' accuracy for ANN over other methods for morbidity (i.e. morbidity has not been sufficient precisely defined in the models, so spurious accuracy has resulted). Thirdly, our study did not have external validation; hence we were not able to suggest that an ANN model developed at one institution can perform as well at another institution. Further studies should be done in different populations to externally validate our findings. In addition, geographic factors or population characteristics can be included in the ANN models to see if they can be universally applied in different hospitals or countries. Nevertheless, the ANN can be developed for local use at any institution and can be usefully integrated with the hospital information system for clinical use.

In summary, risk prediction models are essential for risk assessment, cost-benefit analysis, comparison of quality between institutions and individuals, and study of therapy trends [6, 28–30]. The ANN appears to be very suitable for these purposes. Finally, it should be remembered that the predicted probability, as calculated with these models, only reflects the likelihood of adverse events with average care by an average surgeon, and must to be interpreted cautiously.

References

1 Favaloro RG. Critical analysis of coronary artery bypass grafting: a 30 year journey. *Journal of the American College of Cardiology* 1998; **31**: 1B–63B.

2 Pocock SJ, Henderson RA, Rickards AF, et al. Meta-analysis of randomised trials comparing coronary angioplasty with bypass surgery. *Lancet* 1995; **346**: 1184–9.

3 Disch DL, O'Connor GT, Birkmeyer JD, Olmstead EM, Levy DG, Plume SK. Changes in patients undergoing

coronary artery bypass grafting: 1987-1990. Northern New England Cardiovascular Disease Study Group. *The Annals of Thoracic Surgery* 1994; **57**: 416–23.
4 Estafanous FG, Loop FD, Higgins TL, et al. Increased risk and decreased morbidity of coronary artery bypass grafting between 1986 and 1994. *The Annals of Thoracic Surgery* 1998; **65**: 383–9.
5 Naunheim KS, Fiore AC, Wadley JJ, et al. The changing profile of the patient undergoing coronary artery bypass surgery. *Journal of the American College of Cardiology* 1988; **11**: 494–8.
6 Parsonnet V, Dean D, Bernstein AD. A method of uniform stratification of risk for evaluating the results of surgery in acquired adult heart disease. *Circulation* 1989; **79**: I3–12.
7 Hannan EL, Kilburn H, O'Donnell JF, Lukacik G, Shields EP. Adult open heart surgery in New York State. An analysis of risk factors and hospital mortality rates. *Journal of the American Medical Association* 1990; **264**: 2768–74.
8 Higgins TL, Estafanous FG, Loop FD, Beck GJ, Blum JM, Paranandi L. Stratification of morbidity and mortality outcome by preoperative risk factors in coronary artery bypass patients. A clinical severity score. *Journal of the American Medical Association* 1992; **267**: 2344–8.
9 O'Connor GT, Plume SK, Olmstead EM, et al. Multivariate prediction of in-hospital mortality associated with coronary artery bypass graft surgery. Northern New England Cardiovascular Disease Study Group. *Circulation* 1992; **85**: 2110–8.
10 Tremblay NA, Hardy JF, Perrault J, Carrier M. A simple classification of the risk in cardiac surgery: the first decade. *Canadian Journal of Anaesthesia* 1993; **40**: 103–11.
11 Tu JV, Jaglal SB, Naylor CD. Multicenter validation of a risk index for mortality, intensive care unit stay, and overall hospital length of stay after cardiac surgery. Steering Committee of the Provincial Adult Cardiac Care Network of Ontario. *Circulation* 1995; **91**: 677–84.
12 Tuman KJ, McCarthy RJ, March RJ, Najafi H, Ivankovich AD. Morbidity and duration of ICU stay after cardiac surgery. A model for preoperative risk assessment. *Chest* 1992; **102**: 36–44.
13 Higgins TL. Quantifying risk and assessing outcome in cardiac surgery. *Journal of Cardiothoracic and Vascular Anesthesia* 1998; **12**: 330–40.
14 Dupuis JY, Wang F, Nathan H, Lam M, Grimes S, Bourke M. The cardiac anesthesia risk evaluation score: a clinical useful predictor of mortality and morbidity after cardiac surgery. *Anesthesiology* 2001; **94**: 194–204.
15 Chiu JS, Li YC, Yu FC, Wang YF. Applying an artificial neural network to predict osteoporosis in the elderly. *Studies in Health Technology and Informatics* 2006; **124**: 609–14.
16 Peng SY, Wu KC, Wang JJ, Chuang JH, Peng SK, Lai YH. Predicting postoperative nausea and vomiting with the application of an artificial neural network. *British Journal of Anaesthesia* 2007; **98**: 60–5.
17 Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network classification models: a methodology review. *Journal of Biomedical Informatics* 2002; **35**: 352–9.
18 Kantardzic M. *Data Mining*. New Jersey: Wiley-IEEE Press, 2002.
19 Heijmans JH, Maessen JG, Roekaerts PM. Risk stratification for adverse outcome in cardiac surgery. *European Journal of Anaesthesiology* 2003; **20**: 515–27.
20 Nashef SA, Roques F, Michel P, et al. European system for cardiac operative risk evaluation (EuroSCORE). *European Journal of Cardio-Thoracic Surgery* 1999; **16**: 9–13.
21 Buzatu DA, Taylor KK, Peret DC, Darsey JA, Lang NP. The determination of cardiac surgical risk using artificial neural networks. *The Journal of Surgical Research* 2001; **95**: 61–6.
22 Ennett CM, Frize M. Weight-elimination neural networks applied to coronary surgery mortality prediction. *IEEE Transactions on Information Technology in Biomedicine* 2003; **7**: 86–92.
23 Lippmann RP, Shahian DM. Coronary artery bypass risk prediction using neural networks. *The Annals of Thoracic Surgery* 1997; **63**: 1635–43.
24 Orr RK. Use of a probabilistic neural network to estimate the risk of mortality after cardiac surgery. *Medical Decision Making* 1997; **17**: 178–85.
25 Tu JV, Weinstein MC, McNeil BJ, Naylor CD. Predicting mortality after coronary artery bypass surgery: what do artificial neural networks learn? The Steering Committee of the Cardiac Care Network of Ontario. *Medical Decision Making* 1998; **18**: 229–35.
26 Grover FL, Hammermeister KE, Shroyer AL. Quality initiatives and the power of the database. What they are and how they run. *The Annals of Thoracic Surgery* 1995; **60**: 1514–21.
27 Jones RH, Hannan EL, Hammermeister KE, et al. Identification of preoperative variables needed for risk adjustement of short-term mortality after coronary artery bypass graft surgery. The Working Group Panel on the Cooperative CABG Project. *Journal of the American College of Cardiology* 1996; **28**: 1478–87.
28 Edwards FH, Albus RA, Zajtchuk R, Graeber GM, Barry M. A quality assurance model of operative mortality in coronary artery surgery. *The Annals of Thoracic Surgery* 1989; **47**: 646–9.
29 Griffith BP, Hattler BG, Hardesty RL, Kormos RL, Pham SM, Bahnson HT. The need for accurate risk-adjusted measures of outcome in surgery. Lessons learned through coronary artery bypass. *Annals of Surgery* 1995; **222**: 593–9.
30 Kouchoukos NT, Anderson RP, Fosburg RG, et al. Report of the Ad Hoc Committee on Physician-Specific mortality rates for cardiac surgery. *The Annals of Thoracic Surgery* 1993; **56**: 1200–2.