

# Stratification of Adverse Outcomes by Preoperative Risk Factors in Coronary Artery Bypass Graft Patients: An Artificial Neural Network Prediction Model

Chee-Fah Chong<sup>1,3</sup>, MD, Yu-Chuan Li<sup>1,2</sup>, PhD, Tzong-Luen Wang<sup>3</sup>, PhD, Hang Chang<sup>3</sup>, PhD

<sup>1</sup>Graduate Institute of Medical Informatics, Taipei Medical University

<sup>2</sup>Department of Dermatology, Taipei Medical University Hospital

<sup>3</sup>Emergency Department, Shin Kong Wu Ho-Su Memorial Hospital, Taipei, Taiwan

Correspondence: Yu-Chuan Li, PhD; e-mail: jack@tmu.edu.tw

## ABSTRACT

*We constructed and internally validated an artificial neural network (ANN) model for prediction of in-hospital major adverse outcomes (defined as death, cardiac arrest, coma, renal failure, cerebrovascular accident, reinfarction, or prolonged mechanical ventilation) in patients who received "on-pump" coronary artery bypass grafting (CABG) surgery. We retrospectively analyzed a 5-year CABG surgery database with a final study population of 563 patients. Predictive variables were limited to information available before the procedure, and outcome variables were represented only by events that occurred postoperatively. The ANN's ability to discriminate outcomes was assessed using receiver-operating characteristic (ROC) analysis and the results were compared with a multivariate logistic regression (LR) model and the QMMI risk score (RS) model. A major adverse outcome occurred in 12.3% of all patients and 18 predictive variables were identified by the ANN model. Pairwise comparison showed that the ANN model significantly outperformed the RS model (AUC = 0.886 vs. 0.752,  $p = 0.043$ ). However, the other two pairs, ANN vs. LR models (AUC = 0.886 vs. 0.807,  $p = 0.076$ ) and LR vs. RS models (AUC = 0.807 vs. 0.752,  $p = 0.453$ ) performed similarly well. ANNs tend to outperform regression models and might be a useful screening tool to stratify CABG candidates preoperatively into high-risk and low-risk groups.*

## INTRODUCTION

CABG surgery is one of the most commonly performed major operations in the developed countries. Traditionally, operative mortality has been the primary criterion for judging the quality of surgical results, and several models have been developed to predict mortality based on preoperative data. However, death after CABG surgery is an uncommon event, with an incidence of <4%. The ability to predict operative mortality is important to patients, families, and physicians, but it is an inadequate method of determining surgical outcome<sup>1,2</sup>. The frequency of

major morbidity and mortality after CABG can vary widely across institutions and surgeons. Comparisons of mortality rates without considering morbidity may reach incorrect conclusions regarding quality of care. Major morbidity is more common than mortality after CABG and might have greater economic importance because it results in a prolonged hospital stay and greater utilization of health care resources.

ANNs are computer models composed of parallel, nonlinear computational elements ("neurons") arranged in highly interconnected layers with a structure that mimics biologic neural networks. The ANN can be trained with data from cases that have a known outcome. The network can evaluate the input data, recognize any pattern that may be present, and apply this knowledge to the evaluation of unknown cases. In the analysis of large data sets, ANNs have the advantage of relative insensitivity to noise while having the ability to discover patterns that may not be apparent to human observers<sup>3</sup>. Recently, there has been widespread interest in using ANNs for an extraordinary range of problem domains, in areas as diverse as finance, physics, engineering, geology, and medicine. Anywhere that there are problems of classification, prediction, or control, ANNs are being introduced<sup>4,5,6</sup>.

In this study, we performed a cross-sectional study to develop an ANN model based entirely on preoperative clinical and laboratory data for predicting postoperative major adverse outcomes for patients who received "on-pump" CABG surgery.

## METHODS

### Patient selection and chart review

The study was conducted in Taiwan at Shin Kong Wu Ho-Su Memorial Hospital, which provides urban tertiary care in Taipei city. Patients enrolled in the study included all patients who received a CABG surgery from June 1997 to May 2002. Cases with concomitant surgeries were excluded. "Off-pump" procedures that did not require intraoperative cardiopulmonary bypass

were also excluded. The final cohort, consisting of 563 patients, was randomly divided into 2 mutually exclusive datasets: the derivation set (n = 423) and the validation set (n = 140). The derivation set represents approximately 3/4 of the entire cohort (1/2 for training, 1/4 for cross-verification), and the validation set represents 1/4 of the cohort. The study protocol was approved by the Institutional Review Board (IRB).

### Predictive and predicted variables

An initial literature search was conducted using Medline and other reference sources to identify articles describing factors affecting CABG outcome<sup>7,8</sup>. This initial list of variables was discussed with several cardiac surgeons to determine what variables they believed were most significant in determining the outcome of “on-pump” solitary CABG surgery. Intraoperative surgical characteristics were not considered in the prediction model. Only variables available preoperatively were taken into account. The resulting list of 21 variables included age, gender, history of diabetes mellitus (DM), history of hypertension, prior CABG, history of stroke or transient ischemic attack (TIA), history of uremia or end-stage renal disease, history of severe liver disease, history of peripheral vascular disease, history of chronic obstructive pulmonary disease (COPD), priority of operation, left ventricular ejection fraction, recent percutaneous coronary intervention (PCI) with complication(s), cardiogenic shock, recent myocardial infarction of <1 week, prior endotracheal intubation, prior cardiopulmonary resuscitation (CPR), serum creatinine (Cr), hematocrit, and serum albumin. In-hospital major adverse outcomes included death, cardiac arrest, coma, renal failure (new-onset dialysis), reinfarction, cerebrovascular accident, and prolonged mechanical ventilation of >14 days.

### Data collection

Detailed information of patients in the final cohort was obtained retrospectively after patient discharge through medical record abstraction by one of two researchers. Data collected included demographics, conditions and risks before procedure, last preoperative laboratory results, use of services, postoperative hospital course (adverse events), and discharge laboratory results. For categorical variables, missing data were coded as “missing,” such that for a variable with 3 categories, a fourth was added. For continuous variables, missing data were managed with mean substitution. The average rate of missing information for the 21 independent predictors of adverse outcomes was 2.3%.

### The artificial neural network (ANN) model

Various formulations of a back-propagation ANN were used to train and predict any major adverse outcome

after CABG from 21 independent predictive variables. The neural networks were generated using the software package *STATISTICA Neural Networks (Release 4.0E)* from *StatSoft Inc.* To train each ANN to identify the clinical patterns corresponding to the CABG outcomes, the data were randomly divided into 3 separate sets: a set for training the network, a set for cross verifying it, and a set for validating it. A standard technique in neural networks is to train the network using one set of data, but to check performance against a verification set not used in training: this provides an independent check that the network is actually learning something useful. Without cross verification, a network with a large number of weights and a modest amount of training data can overfit the training data - learning the noise present in the data rather than the underlying structure. This ability of a network not only to learn the training data, but to perform well on previously-unseen data, is known as generalization.

The final best ANN model used in this study was a feed-forward, multilayer perceptron (MLP) network architecture using the back-propagation supervised training algorithm. Neural computation was performed on an IBM compatible Pentium III computer running at 1 GHz. The ANN had an input layer with 18 nodes, a hidden layer with 8 nodes, and an output layer with one node. The ANN topology is depicted in Figure 1. When the root mean square (RMS) error no longer decreased significantly with training, that network was tested by calculating the classification accuracy on the validation set. Once the best network was identified, a pruning procedure was performed in which those processing elements that did not contribute significantly to the overall predictive power of the network were removed. The performance of the final ANN was then evaluated by calculating the area under the ROC curve (AUC), and the results compared with the other two models.

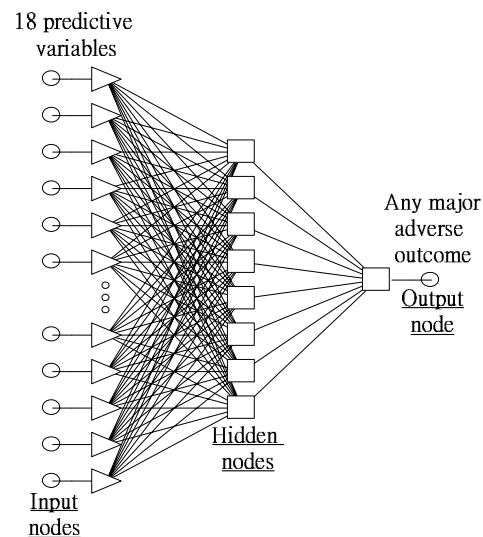


Figure 1. Diagram of an (18 x 8 x 1) ANN model.

### The multivariate logistic regression (LR) model

A multivariate logistic regression (LR) model was developed utilizing *SPSS for Windows 10.0 (SPSS Inc)*. The model was constructed on the derivation set (n = 463, with 10 cases containing missing data). No variable selection procedure was applied. All the variables unconditionally entered the logistic regression equation because they were already deemed significant in the second step of the variable selection process. Its performance was assessed with ROC analysis using the same validation set and the results compared with the other two models.

### The risk score (RS) model

We applied a clinical prediction rule developed by Fortescue et al<sup>9</sup> using multivariate logistic regression analysis to evaluate its performance on the same validation set of our study cohort. This so called Quality Measurement and Management Initiative (QMMI) score included sixteen independent predictive variables of major adverse outcomes (Table 1). Its performance was then compared with that of the other

two models.

**Table 1. The QMMI score.**

Correlate	OR	Confidence Intervals	Points
Cr $\geq$ 3.0 mg/dL	4.4	2.8–7.0	12
Age $\geq$ 80 yrs	4.0	2.6–6.1	11
Cardiogenic shock	3.2	1.9–5.4	10
Emergent operation	2.8	2.0–4.0	9
Age 70–79 yrs	2.5	1.8–3.4	8
Prior CABG	2.3	1.7–3.1	7
Ejection fraction <30%	2.1	1.5–3.1	6
Liver disease (history)	2.1	1.1–4.1	6
Age 60–69 yrs	1.9	1.4–2.6	5
Cr 1.5–3.0 mg/dL	1.9	1.4–2.4	5
Stroke or TIA (history)	1.6	1.2–2.1	4
EF 30%–49%	1.4	1.1–2.0	3
COPD (history)	1.5	1.1–1.8	3
Female gender	1.4	1.1–1.8	3
Hypertension (history)	1.3	1.0–1.6	2
Urgent operation	1.3	1.0–1.6	2

From Fortescue et al<sup>9</sup>: *Am J Cardiol*. 2001;88:1251-8.

**Table 2. Patient characteristics: comparison of derivation and validation sets.**

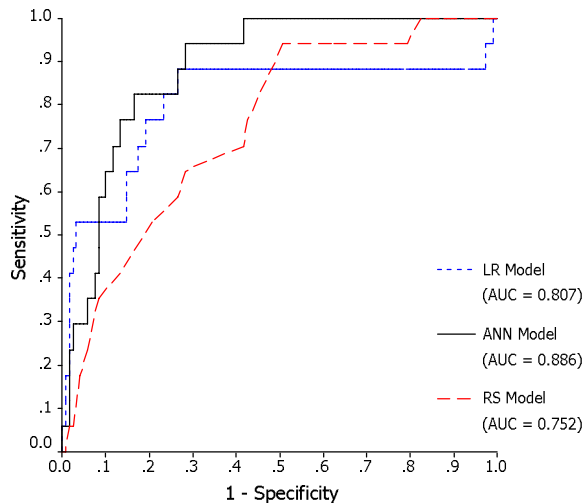
Characteristic	Derivation (n = 423)	Validation (n = 140)
<b>Dichotomous Variables: n (%)</b>		
Female gender	128 (30.3)	37 (26.4)
Diabetes Mellitus	154 (36.3)	57 (40.4)
Hypertension	251 (59.3)	69 (49.3)
Prior CABG surgery	3 (0.7)	1 (0.7)
Uremia or end-stage renal disease	12 (2.8)	3 (2.1)
Stroke or transient ischemic attack	26 (6.1)	9 (6.4)
Peripheral vascular disease	13 (3.1)	4 (2.9)
Liver disease*	1 (0.2)	0 (0)
COPD	12 (2.8)	3 (2.1)
Complicated PCI	8 (1.9)	3 (2.1)
Cardiogenic shock	10 (2.4)	2 (1.4)
Prior endotracheal intubation*	2 (0.5)	0 (0)
Prior CPR*	1 (0.2)	0 (0)
Recent myocardial infarction	28 (6.6)	10 (7.1)
<b>Categorical Variables: n (%)</b>		
Priority of surgery		
• Emergent	7 (1.7)	2 (1.4)
• Urgent	7 (1.7)	3 (2.1)
• Elective	409 (96.7)	135 (96.4)
Left ventricular ejection fraction		
• $\leq$ 30%	11 (2.6)	2 (1.4)
• 31–40%	36 (8.5)	15 (10.7)
• 41–50%	32 (7.6)	15 (10.7)
• >50%	344 (81.3)	108 (77.1)
Coronary artery stenosis		
• Left main disease	118 (27.9)	36 (25.7)
• Three-vessel disease	276 (65.2)	98 (70.0)
• Others	29 (6.9)	6 (4.3)
<b>Continuous Variables: mean (SD)</b>		
Age (yr)	64.0 (10.4)	63.3 (9.7)
Serum creatinine (mg/dL)	1.4 (1.3)	1.5 (1.6)
Hematocrit (%)	37.7 (5.5)	37.7 (4.7)
Serum albumin (g/dL)	4.0 (1.5)	3.9 (0.5)

## Comparison of models

ROC curves of the validation set were plotted for the three models (ANN, LR, and RS) and the AUCs calculated for comparison. The ROC curve plots the false-positive rate (1 - specificity) on the x-axis and the true-positive rate (sensitivity) on the y-axis. The area under the ROC curve is a measure of a model's discriminatory power. The larger the area under the ROC curve, the more accurate the prediction model. The sensitivity and specificity at a cut-off value corresponding with the highest accuracy (minimal false negative and false positive results) were also computed and compared.

## RESULTS

The final study population included 563 patients who underwent "on-pump" CABG not involving any concomitant procedures, with 423 patients allocated at random to the derivation set and the remaining 140 to the validation set (Table 2). Differences in preoperative patient characteristics were not significant across groups. One or more major adverse outcomes were found to occur in 12.3% of patients (10 cases in the derivation set and 3 in the validation set). Of these there were 42 deaths (7.5%), 30 episodes of prolonged mechanical ventilation (5.3%), 23 episodes of renal failure (4.1%), 14 strokes (2.5%), 12 episodes of cardiac arrest (2.1%), 9 episodes of reinfarction (1.6%), and 4 episodes of coma (0.7%).



**Figure 2. Comparison of ROC curves (validation set).**

Figure 2 depicts the ROC curves for the ANN, LR, and RS models. Results of pairwise comparison of the ROC curves were shown in Table 3. The ANN model significantly outperformed the RS model (AUC = 0.886 vs. 0.752,  $p = 0.043$ ). However, the other two pairs,

ANN vs. LR models (AUC = 0.886 vs. 0.807,  $p = 0.076$ ) and LR vs. RS models (AUC = 0.807 vs. 0.752,  $p = 0.453$ ) performed similarly well.

**Table 3. Pairwise comparison of ROC curves (validation set).**

Pairwise comparison	Standard Error	95% CI		P value
		Lower	Upper	
ANN vs. LR Models	0.044	-0.008	0.165	0.076
ANN vs. RS Model	0.066	0.004	0.263	<b>0.043</b>
LR vs. RS Model	0.073	-0.089	0.199	0.453

If a cut-off value corresponding with minimal false negative and false positive results was selected, the sensitivity and specificity of each model can be calculated and compared. (Table 4).

**Table 4. Testing summary on the validation set assuming highest accuracy.**

	ANN Model	LR Model	RS Model
<b>Sensitivity</b>	94.1 %	88.2 %	76.5 %
<b>Specificity</b>	71.7 %	73.3 %	57.5 %

## DISCUSSION

Although several severity scores exist that attempt to predict outcomes after CABG surgery, most of them are specific for mortality prediction. Few published studies to date have attempted to develop a model that predicts major adverse outcomes including mortality<sup>9,10,11</sup>. In our study, we constructed and internally validated an ANN model that estimated the risk of major adverse outcomes after CABG surgery. A neural computational severity model such as this has the potential to meet several important roles, such as assisting clinical decision making, facilitating more accurate comparisons of outcomes across groups, and evaluating trends in cardiac surgical practices over time.

The area under the ROC curve (AUC) is a measure of a model's discriminatory power. According to the observation by Swets et al<sup>12</sup>, an AUC of  $\geq 0.7$  is diagnostically useful. In our study, all three models discriminate well (AUC  $\geq 0.7$ ). The ANN model significantly outperforms the risk score model (AUC = 0.886 vs. 0.752,  $p = 0.043$ ) in the validation set. The ANN model also has better simultaneous sensitivity and specificity at the cut-off value corresponding to highest accuracy. It seems that the ANN model has a higher predictive power than the QMMI score.

Pairwise comparison of performance between the LR and RS models (AUC = 0.807 vs. 0.752) showed no significant difference ( $p = 0.453$ ) since both models are based on similar statistics: logistic regression analysis. Though the performance of ANN (AUC = 0.886) is better than that of the LR model (AUC = 0.807), it is statistically insignificant ( $p = 0.076$ ). However, the comparison between ANN performance and logistic regression is not perfect because patients with missing data will be excluded in logistic regression, whereas the network will test all patients regardless of missing data.



At a cut-off value corresponding with highest accuracy, the ANN tends to have the best simultaneous sensitivity and specificity among the 3 models. We believe that, if a larger dataset containing more missing data is used, an ANN model should have better performance than logistic regression or risk score model.

All prior studies using regression methodologies to predict outcome of CABG have been based on the assumption that all required patient information would be available at the time of initial evaluation. Frequently, information is not available, either because patients can be poor historians or the practitioner forgets to ask pertinent questions, forgets to collect key observations, or is unable to retrieve all laboratory results in a timely fashion. In this study cohort, an average of 2.3% of the data was missing on all patients. Yet, the ANN accuracy was not significantly compromised by this absence.

A neural network approach is also preferable in that ANNs are model independent and very flexible in being able to use mixes of categorical and continuous variables. ANNs have the added advantage that they can learn to predict arbitrarily complex nonlinear relationships between independent and dependent variables by including more processing elements in the hidden layer or more hidden layers in the network. These advantages make the ANN a more robust paradigm for application to a real-world setting<sup>13,14</sup>. The use of the network in real time will not be difficult. The number of hospitals that have or will install an electronic medical record is growing rapidly. Once trained, the network could reside in the background of the clinical information systems. The data used by the network represent the standard information routinely collected on CABG candidates admitting to the wards or intensive care units. Once entered into the electronic record, these data could then be parsed and used by the network to generate a probability of the predicted outcome. Network accuracy could also be continuously improved over time because it can constantly be retrained as more patients are accumulated.

On the contrary, the "black box" interpretation is a major obstacle to the acceptance of ANNs as one mechanism for the medical decision support systems. However, an accurate second opinion is often helpful in medical decision making with or without a detailed understanding of how it works.

There are a number of limitations to this study that need to be addressed. First, the ANN was not tested in real time. It is not clear how physicians or surgeons will respond if given network predicted outcome prior to surgery. Second, patients with severe liver disease and episodes of prior endotracheal intubation or prior CPR were so rare that they were "pruned" during the training process. Finally, this study was carried out at a single institution. These findings must be corroborated on patients from multiple locations. Portability will be a critical factor to the future use of the ANN in this setting.

The ANN model tends to outperform regression models in the prediction of postoperative major adverse

outcomes for CABG patients and it might serve as a useful screening tool to stratify CABG candidates preoperatively into high-risk and low-risk groups. Physicians could then use this information to aid them in making their treatment and final disposition.

## REFERENCES

1. Green J, Wintfeld N, Sharkey P, Passman LJ. The importance of severity of illness in assessing hospital mortality. *JAMA* 1990;263:241-246.
2. Silber JH, Rosenbaum PR, Schwartz JS, Ross RN, Williams SV. Evaluation of the complication rate as a measure of quality of care in coronary artery bypass graft surgery. *JAMA* 1995;274:317-323.
3. Penny W, Frost D. Neural networks in clinical medicine. *Med Decis Making*. 1996;16:386-98.
4. Baxt WG, Skora J. Prospective validation of artificial neural network trained to identify acute myocardial infarction. *Lancet*. 1996;347:12-5.
5. Resnic FS, Popma JJ, Ohno-Machado L. Development and evaluation of models to predict death and myocardial infarction following coronary angioplasty and stenting. *Proc AMIA Symp*. 2000;:690-3.
6. Li YC, Liu L, Chiu WT, Jian WS. Neural network modeling for surgical decisions on traumatic brain injury patients. *Int J Med Inf*. 2000;57:1-9.
7. Jones RH, Hannan EL, Hammermeister KE, DeLong ER, O'Connor GT, Luepker RV, Parsonnet V, Pryor DB. Identification of preoperative variables needed for risk adjustment of short-term mortality after coronary artery bypass graft surgery. The Working Group Panel on the Cooperative CABG Database Project. *J Am Coll Cardiol*. 1996;28:1478-87.
8. Tu JV, Sykora K, Naylor CD. Assessing the outcomes of coronary artery bypass graft surgery: how many risk factors are enough? Steering Committee of the Cardiac Care Network of Ontario. *J Am Coll Cardiol*. 1997;30:1317-23.
9. Fortescue EB, Kahn K, Bates DW. Development and validation of a clinical prediction rule for major adverse outcomes in coronary bypass grafting. *Am J Cardiol*. 2001;88:1251-8.
10. Higgins TL, Estafanous FG, Loop FD, Beck GJ, Blum JM, Parandhi L. Stratification of morbidity and mortality outcome by preoperative risk factors in coronary artery bypass patients: A clinical severity score. *JAMA* 1992;267:2344-2348.
11. Magovern JA, Sakert T, Magovern GJ, Benckart DH, Burkholder JA, Liebler GA, et al. A model that predicts morbidity and mortality after coronary artery bypass graft surgery. *J Am Coll Cardiol* 1996;28:1147-1153.
12. Swets JA. Measuring the accuracy of diagnostic systems. *Science* 1988;240:1285-1293.
13. Tu JV. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *J Clin Epidemiol* 1996;49:1225-1231.
14. Lin CS, Li YC, Mok MS, Wu CC, Chiu HW, Lin YH. Neural network modeling to predict the hypnotic effect of propofol bolus induction. *Proc AMIA Symp*. 2002;:450-4.