Risk factor identification and mortality prediction in cardiac surgery using artificial neural networks

Johan Nilsson, MD, PhD,[a] Mattias Ohlsson, PhD,[b] Lars Thulin, MD, PhD,[a] Peter Höglund, MD, PhD,[c] Samer A.M. Nashef, FRCS,[d] and Johan Brandt, MD, PhD[a]

See related editorial on page 8.

Supplemental material is available online.

Objective: The artificial neural network model is a nonlinear technology useful for complex pattern recognition problems. This study aimed to develop a method to select risk variables and predict mortality after cardiac surgery by using artificial neural networks.

Methods: Prospectively collected data from 18,362 patients undergoing cardiac surgery at 128 European institutions in 1995 (the European System for Cardiac Operative Risk Evaluation database) were used. Models to predict the operative mortality were constructed using artificial neural networks. For calibration a sixfold cross-validation technique was used, and for testing a fourfold cross-testing was performed. Risk variables were ranked and minimized in number by calibrated artificial neural networks. Mortality prediction with 95% confidence limits for each patient was obtained by the bootstrap technique. The area under the receiver operating characteristics curve was used as a quantitative measure of the ability to distinguish between survivors and nonsurvivors. Subgroup analysis of surgical operation categories was performed. The results were compared with those from logistic European System for Cardiac Operative Risk Evaluation analysis.

Results: The operative mortality was 4.9%. Artificial neural networks selected 34 of the total 72 risk variables as relevant for mortality prediction. The receiver operating characteristics area for artificial neural networks (0.81) was larger than the logistic European System for Cardiac Operative Risk Evaluation model (0.79; $P = .0001$). For different surgical operation categories, there were no differences in the discriminatory power for the artificial neural networks ($P = .15$) but significant differences were found for the logistic European System for Cardiac Operative Risk Evaluation ($P = .0072$).

Conclusions: Risk factors in a ranked order contributing to the mortality prediction were identified. A minimal set of risk variables achieving a superior mortality prediction was defined. The artificial neural network model is applicable independent of the cardiac surgical procedure.

From the Departments of Cardiothoracic Surgery[a] and Theoretical Physics,[b] Lund University; Competence Centre for Clinical Research,[c] Lund University Hospital, Lund, Sweden; and Papworth Hospital,[d] Cambridge, UK.

Supported by grants from the Swedish Heart Lung Foundation, Anna Lisa and Sven-Eric Lundgrens foundation for medical research, and by computer resources from Lunarc at Lund University.

Received for publication Sept 1, 2005; revisions received Dec 19, 2005; accepted for publication Dec 29, 2005.

Address for reprints: Dr Johan Nilsson, Department of Cardiothoracic Surgery, Heart and Lung Center, Lund University Hospital, SE 221 85 LUND, Sweden (E-mail: johan.nilsson@thorax.lu.se).

J Thorac Cardiovasc Surg 2006;132:12-9

0022-5223/$32.00

doi:10.1016/j.jtcvs.2005.12.055

Preoperative evaluation of a patient's surgical risk is an important component in cardiac surgery. Risk stratification can provide patients and their families with insight into the existent risk of complications and mortality and guide the selection of cases for surgery versus alternative, nonsurgical therapies. It can

Abbreviations and Acronyms

ANN = artificial neural network
CABG = coronary artery bypass grafting
CI = confidence interval
ROC = receiver operating characteristics
SD = standard deviation

also predict the need for hospital care resources in cardiac surgery.[1] During the last decades, several scoring systems to calculate the mortality risk before the surgery have been developed.[2-5]

Most risk scoring systems have been created using a biostatistical method based on a generalized linear model with assumptions of linear relationship. Artificial neural networks (ANNs) work in a nonlinear fashion, which may better describe the interaction between health risk factors. ANNs have been used in classification and diagnostic prediction of cancer[6] and electrocardiogram interpretation,[7] among others. Some studies in clinical medicine have demonstrated superiority of the classification or prediction by ANNs compared with other statistical models.[8] In the field of cardiac surgery, only a few studies using ANNs have been published, and the results have been ambiguous.[9-14]

To select risk variables for a model, significance testing (*P* values) is the most common methodology, but this does not assess the importance of the individual variable.[15] On the other hand, ANNs may be used for both variable selection and ranking of individual variables in order of importance.[15] For example, this methodology has been employed to select and minimize a large number of gene expression levels used in cancer classification, with excellent results.[16]

This study aimed to systematically evaluate the accuracy and performance of ANNs to select and rank the most important risk factors for operative mortality in cardiac surgery by using high-performance computer clusters.

Methods

Database

The database used in the present study was that of the multinational European System for Cardiac Operative Risk Evaluation (EuroSCORE) cardiac surgical project. This was a prospective study to assess risk factors for operative mortality, defined as death within 30 days after the operation or within the same hospital admission,[17] and to construct a risk stratification system.[5] The database included 97 risk factors from all patients who underwent cardiac surgery in 128 centers from 8 European countries from September to December 1995. The data collection, quality checks, and validation have been described by Roques and colleagues.[17] A local database including risk factors for adult patients undergoing cardiac surgery at the Lund University Hospital between January 1996 and February 2001 was used to further evaluate the developed ANN risk model by blind testing.

Patients and Study Design

From the 97 original EuroSCORE variables, a subset of 72 variables was selected (Tables 1 and 2). This was done by excluding variables closely linked to other variables and data collected intraoperatively (ie, number of conduits and number of distal coronary anastomoses). Patients with a missing value in any mandatory variable (age, gender, or surgical procedure) or outcome (operative mortality) were excluded from analysis. Imputation was used to substitute missing values in the other variables with the statistical mode for categorical variables and the mean for continuous variables.[11]

Calibration of the ANN Model and Selection of Risk Factors Utilized for Mortality Prediction

An ensemble approach was used where several ANNs were combined into a single prediction model. The individual members of the ensemble were standard multilayer perceptrons with 1 hidden layer and 1 output node that was used to encode the operative mortality.[18] The model selection was performed using a sixfold cross-validation procedure (Figure 1). To select the most important risk variables and to minimize the number of variables included in the final model, a ranking of risk variables was performed.[15]

Performance and Accuracy

The performance and accuracy of the ANN model was compared with the logistic EuroSCORE model[19] and a logistic model. The final prediction models were tested on patients not previously exposed to the models by using a fourfold cross-testing technique (Figure 1).

Statistical Analysis

Mean values (± standard deviation [SD]) were used to describe continuous variables, and frequencies were calculated for categorical variables. Logistic regression analysis was performed to obtain the coefficients for the risk variables included in the logistic model as described by Hosmer and Lemeshow.[20]

To compare the number of correctly classified patients by ANNs versus the logistic EuroSCORE, a proportion test was used. Effective odds ratio for the risk variables were determined as described by Lippmann and Shahian.[11] The 95% confidence intervals (CIs) for both the odds ratio and the output from the ANNs were calculated using the bootstrap technique.[11,21]

Receiver operating characteristics (ROC) curves were used to describe the performance and predictive accuracy for the models.[22] The area (with 95% CI) under the curve was used as a quantitative measure of the ability of the risk predictor models to distinguish between survivors and nonsurvivors. To compare the areas under the resulting ROC curves, the nonparametric approach described by DeLong and coworkers[23] was used.

Computer Cluster and Software

High-performance computing clusters were used to train and evaluate the ANNs. The ANN calibration and analyses were performed with MatLab 7 (2005), Neural Network Toolbox (MathWorks, Natick, Mass). Graphs and statistical analyses were performed using the Intercooled Stata version 9.0 (2005) statistical package (StataCorp LP, College Station, Tex).

TABLE 1. Preoperative risk factors relevant for the prediction of operative mortality, ranked in order of influence upon discriminatory power*

Rank no.	Risk variable	Mean (SD) or n (%)	Odds ratio (95% CI)
1	Age (y)	62.6 (10.7)	1.042 (1.037-1.047)
2	One previous cardiac operation	1137 (6.2)	3.001 (2.645-3.385)
3	Left ventricular ejection fraction	55.5 (14.8)	0.985 (0.981-0.988)
4	Serum creatinine (μmol/L)	103.5 (49.2)	1.004 (1.003-1.005)
5	Emergency operation (<24 h)	893 (4.9)	3.258 (2.839-3.769)
6	Acute aortic dissection	159 (0.9)	7.902 (6.069-10.280)
7	Thoracic aortic surgery	295 (1.6)	6.316 (5.275-7.538)
8	Heart or heart-lung transplantation	129 (0.7)	5.462 (3.673-7.603)
9	Aortic valve surgery for stenosis	2655 (14.5)	1.529 (1.349-1.743)
10	Acute active endocarditis	192 (1.0)	4.72 (3.624-6.042)
11	Urgent operation (stay in hospital before surgery)	3775 (20.6)	1.594 (1.443-1.763)
12	Mitral valve surgery for stenosis	946 (5.2)	2.805 (2.201-3.536)
13	Chronic congestive heart failure	1787 (9.7)	1.926 (1.712-2.172)
14	Intubated (before arrival in the operating room)	194 (1.1)	7.598 (5.808-9.846)
15	Carotid disease (unilateral stenosis >50%)	301 (1.6)	1.928 (1.439-2.454)
16	Intravenous inotropic support	425 (2.3)	5.727 (4.180-7.708)
17	Coronary bypass grafting	13286 (72.4)	1.305 (1.127-1.505)
18	Patient refusal of blood products	44 (0.2)	0.308 (0.257-0.375)
19	Atrial fibrillation	1676 (9.1)	1.392 (1.177-1.625)
20	Height (cm)	167.9 (9.0)	0.983 (0.978-0.988)
21	Hematocrit (%)	39.9 (4.8)	1.009 (1.000-1.019)
22	Long-term immunosuppressive therapy	76 (0.4)	4.249 (3.087-5.612)
23	Pulmonary embolectomy	14 (0.1)	18.161 (6.681-37.427)
24	Intra-aortic balloon pump	184 (1.0)	3.93 (2.943-5.185)
25	Previous surgery for vascular disease (carotids)	166 (0.9)	2.858 (2.013-3.757)
26	Intermittent claudication	1088 (5.9)	2.159 (1.926-2.425)
27	Systolic pulmonary artery pressure >60 mm Hg	361 (2.0)	3.05 (2.484-3.750)
28	Tricuspid valve surgery	309 (1.7)	3.956 (3.115-4.997)
29	Postinfarction ventricular septal rupture closure	39 (0.2)	7.093 (4.892-10.326)
30	Neurologic disorder	257 (1.4)	2.653 (2.146-3.205)
31	Cardiogenic shock	532 (2.9)	2.265 (1.638-3.064)
32	Mitral surgery for ischemic acute regurgitation	49 (0.3)	3.168 (1.619-4.821)
33	No ITA (preoperative decision)	1480 (8.1)	1.799 (1.607-2.002)
34	Recent myocardial infarction (number of days ago)	34.7 (25.2)	0.996 (0.994-0.998)

ITA, Internal thoracic artery. *See Figure 2, A.

Results

Patient Population

From the EuroSCORE database of 19,030 patients, 668 patients were excluded due to a missing value in any of the mandatory variables (age, gender, or surgical procedure) or outcome (operative mortality). Thus, 18,362 patients were included in the analysis. In 0.7% of the total data points missing values were imputed, as previously described.

The average age was 62.6 ± 10.7 years (range 17-89). The majority of patients were men (72%). Isolated coronary artery bypass grafting (CABG) was performed in 11,628 patients (63%), 4907 (27%) patients had a valve procedure with or without CABG surgery, and 1827 (10%) had miscellaneous procedures. The patient details are described in Table 1. The actual operative mortality was 4.9% (n = 891).

Architecture of Artificial Neural Networks

Approximately 42,500 different ANN models were validated. The architecture for the final validation ANN included 1 hidden layer with 14 nodes, 1 output node, and 6 individual members of the ensemble. This ANN architecture was used in the selection of risk factors utilized for the mortality prediction.

Selection of Risk Factors Utilized for Mortality Prediction

The importance ranking order of the risk variables for the ANN model is presented in Tables 1 and 2 and Figure 2, *A*. To optimize the model, an increasing number of the ranked variables was included in the model, starting with the top-ranked variable. The largest validation ROC area, 0.82

TABLE 2. Preoperative risk factors with no or negative influence on the prediction of operative mortality, ranked in order of influence upon discriminatory power*

Rank no.	Risk variable	Mean (SD) or n (%)
35	Female gender	5194 (28.3)
36	Past chronic renal failure (no dialysis)	539 (2.9)
37	Two previous cardiac operations	141 (0.8)
38	Left ventricular aneurysmectomy	125 (0.7)
39	Past chronic renal failure (dialysis)	106 (0.6)
40	Diastolic blood pressure (mm Hg)	75.7 (12.3)
41	Angina at rest	2585 (14.1)
42	Carotid disease (bilateral stenosis >50%)	509 (2.8)
43	Ventricular tachycardia/fibrillation	208 (1.1)
44	Angina following recent myocardial infarction	1452 (7.9)
45	Aortic valve surgery for regurgitation	1687 (9.2)
46	More than 2 previous cardiac operations	52 (0.3)
47	Cardiac massage (preoperative)	90 (0.5)
48	Unstable angina (requiring intravenous nitrates)	1495 (8.1)
49	Systolic blood pressure (mm Hg)	132.5 (20.3)
50	Diabetes (oral therapy)	1580 (8.6)
51	Active AIDS (excluding HIV-positive alone)	4 (0.0)
52	Atrial septal defect closure	211 (1.1)
53	Previous surgery for vascular disease (limb arteries)	285 (1.6)
54	Number of diseased coronary vessels	1.7 (1.3)
55	Operation for catheter laboratory complication	182 (1.0)
56	Active neoplasm (malignant tumor known at surgery)	106 (0.6)
57	Mitral valve surgery for regurgitation	1671 (9.1)
58	Urine output <10 mL/h	137 (0.7)
59	Aortic valvular gradient >120 mm Hg	215 (1.2)
60	Diabetes (diet-controlled)	1024 (5.6)
61	Chronic cardiac-related dyspnea at rest	1058 (5.8)
62	Chronic airway disease (treated)	726 (4.0)
63	Weight (kg)	74.4 (13.1)
64	Diabetes (insulin therapy)	719 (3.9)
65	Planned surgery for vascular disease (abdominal aneurysm)	100 (0.5)
66	Permanent pacemaker in place	240 (1.3)
67	Left ventricular aneurysm	231 (1.3)
68	Previous surgery for vascular disease (abdominal aneurysm)	120 (0.7)
69	Planned surgery for vascular disease (limb arteries)	148 (0.8)
70	History of hypertension	8060 (43.9)
71	Left main coronary stenosis (% stenosis)	79.5 (12.1)
72	Planned surgery for vascular disease (carotids)	85 (0.5)

AIDS, Acquired immune-deficiency syndrome; *HIV,* human immune deficiency virus. *See Figure 2, A.

(95% CI: 0.80-0.83), was achieved when the 34 top-ranked risk variables were included (Figure 2, *B*).

To simplify the model, the number of nodes in the hidden layer was decreased until the validation ROC area started to decrease. The optimal ANN finally included 34 risk variables in the input layer and 8 nodes in the hidden layer. All artificial networks from the sixfold cross-validation procedure were saved, resulting in 36 individual networks from the 6 members in the ensemble. Thus, an ensemble size of 36 was used to classify the patients in the test sets.

Performance and Predictive Accuracy for the Algorithms

The discriminatory power (ie, the area under the ROC curve) for operative mortality was significantly larger for the final ANNs, 0.81 (95% CI: 0.79-0.82) compared with the logistic EuroSCORE model, 0.79 (95% CI: 0.78-0.81; $\chi^2 = 15.7$; $P = .0001$; Figure 3). The final ANN ROC area was also significantly larger than the ROC area for a logistic model with the same 34 top-ranked risk variables, 0.80 (95% CI: 0.78-0.81; $\chi^2 = 17.5$, $P < .0001$).

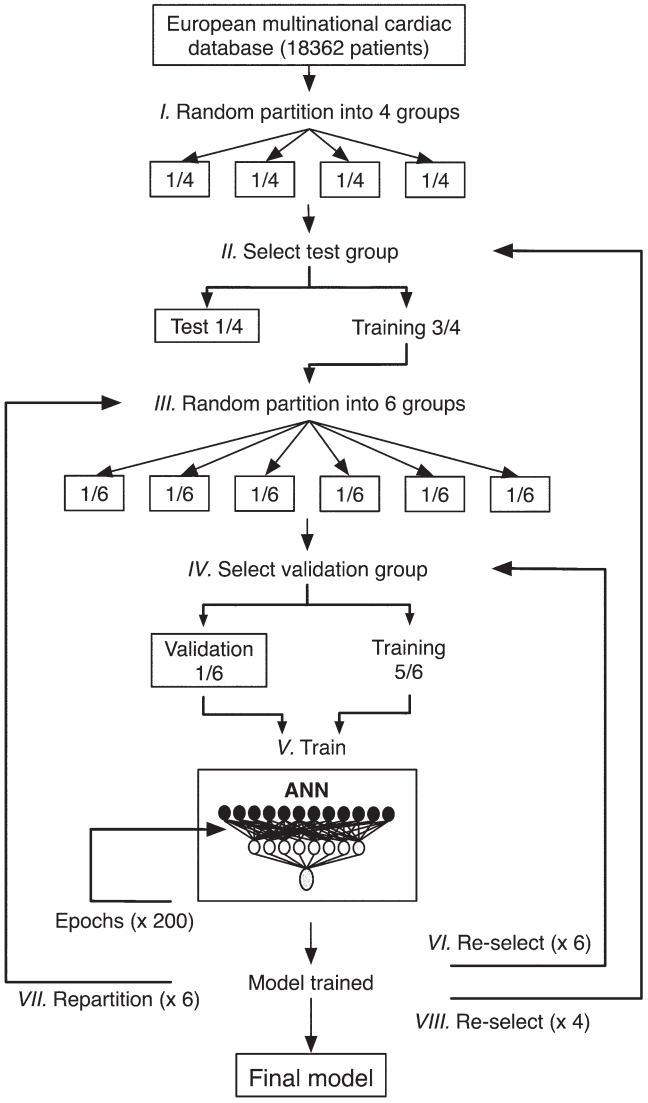

Figure 1. Schematic illustration of the ANN training and analysis process. The cardiac database was randomly split into 4 groups (I). One of these groups was selected as a test set and excluded from further analysis. The remaining groups were used for the training (II). Following the sixfold cross-validation procedure, the training group was randomly partitioned into 6 new groups of equal size (III). One of these groups was reserved for validation and the rest for the actual training (IV). For each model the calibration were optimized with 200 iterations. The procedure was repeated and a new validation group was selected (VI). After 6 reselections all groups had been used for validation, and the training group was repartitioned into six new groups (III) and the entire procedure repeated. After six repartitions (VII) a new test group was selected and the full training process was repeated for the remaining patients (VIII). Thus, for each of the 4 test sets a complete model selection procedure was carried out (steps III-VII) and the final test result was taken as the average over the 4 test sets or by concatenating the 4 test sets into a complete set of test predictions and computing the ROC area for this entire list.

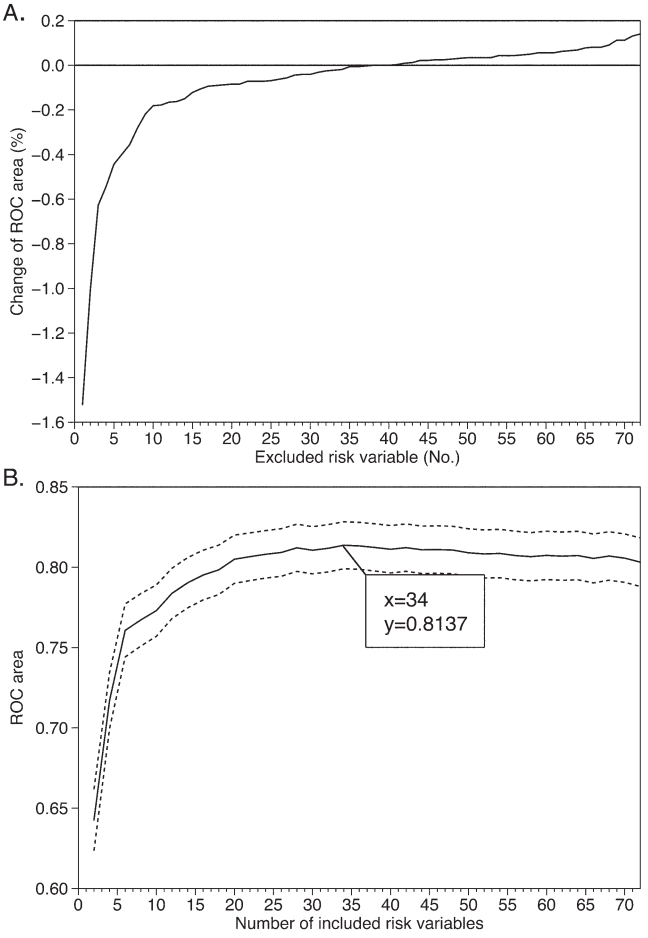

Figure 2. A, The graph shows the difference *(%)* in the validation ROC area *(y-axis)* from the ANN model including 71 risk variables compared with a model including all 72 risk variables. The *x-axis* shows the excluded risk variable number *(No.),* in order of relevance (see Tables 1 and 2). B, The *solid line* shows the validation ROC area *(y-axis)* from the ANNs with different number of included risk variables *(x-axis). Dashed lines* indicate 95% confidence limits.

The numbers of correctly classified survivors for a sensitivity of 25%, 50%, and 75% (corresponding to a ANN mortality risk of 5%, 11%, and 25% and a logistic EuroSCORE mortality risk of 4%, 8%, and 19%) were 17,051, 15,577, and 12,438 patients for the ANNs and 16,990, 15,321, and 11,718 patients for the logistic EuroSCORE. The difference between the ANNs and logistic EuroSCORE was significant for all 3 sensitivity cutoff values ($P = .0395$, $P = .00001$, and $P = .00001$), respectively. For the different surgical procedures there were no differences in discriminatory power for the ANNs, but there were significant differences for the logistic EuroSCORE (Table 3). The discriminatory power was significantly higher for ANNs versus Euro-

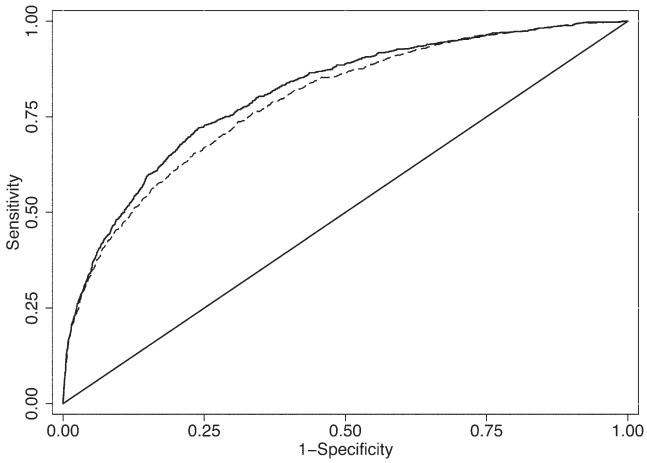

Figure 3. The ROC curves from the test data set. The ANNs *(solid line)* and the logistic EuroSCORE *(dashed line)* risk stratification algorithms. The area under the curve for ANNs is larger compared with the logistic EuroSCORE. $\chi^2 = 15.7$; $P = .0001$.

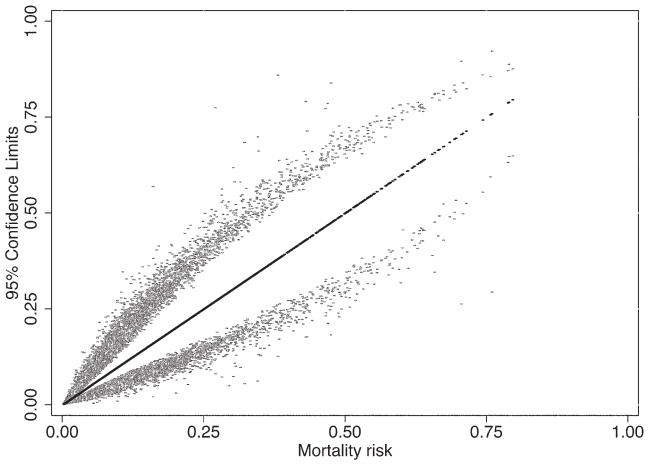

Figure 4. Graph of predicted mortality risk *(x-axis) (black dots)* and the 95% confidence limits *(y-axis) (gray dots)* calculated by the ANNs for each individual patient.

SCORE regarding CABG-only and valve procedures, but the difference did not reach statistical significance in the smaller subset of miscellaneous procedures (Table 3).

Statistical models generally perform best on the population for which they are developed. To evaluate whether the final ANN risk prediction model was applicable in a patient cohort that had not been used in the development of the ANNs or participated in the EuroSCORE project, a subset (n = 1246) from a local database with no missing value in the 34 top-ranked risk variables was used as a blind test. In this cohort the ROC area was 0.83 (95% CI: 0.71-0.94) for the ANNs and 0.80 (95% CI: 0.69-0.90) for the logistic EuroSCORE.

Effective Odds Ratio and Classification Confidence Limits for ANNs

The effective odds ratio for the 34 top-ranked risk variables is presented in Table 1. Bootstrap sampling was used to generate CIs for both the effective odds ratio (Table 1) and the ANN classification (Figure 4). Thus, an individual patient with a calculated mortality risk of 0.64 belongs with 95% certainty to the group of patients with a greater likelihood of not surviving the operation than surviving the surgery. For a patient with a calculated risk of 0.31, the opposite holds true.

Discussion

The purpose of this study was to use ANNs to develop a method for mortality prediction after cardiac surgery and to identify and rank the most important risk factors. The results show that the ANNs had a superior performance and accuracy to predict operative mortality, expressed as discriminatory power and ability to correctly classify patients, compared with the 2 logistic regression models. The ANNs had appropriate power for all cardiac surgical procedures investigated, in contrast to the logistic EuroSCORE model. To our knowledge, this is the first application of ANNs resulting in a significantly superior performance and accuracy to predict operative mortality after cardiac surgery compared with a traditional logistic regression model.

The search for an effective method for mortality prediction in cardiac surgery started in the 1980s.[24] During the last decades, several risk score algorithms for cardiac surgery have been published,[2,4,5,19] but it still remains difficult to

TABLE 3. The ROC area from the test data set for different surgical procedures

Surgical procedure	n	Mortality (%)	ROC area		*P* value
			ANNs	Logistic EuroSCORE	
CABG-only surgery	11,628	371 (3.2)	0.80 (0.77-0.82)	0.78 (0.75-0.80)	.016
Valve procedure*	4907	269 (5.5)	0.76 (0.73-0.79)	0.72 (0.69-0.75)	.0001
Miscellaneous	1827	251 (13.7)	0.80 (0.77-0.83)	0.78 (0.75-0.82)	.072
P value			.15	.0072	

ANNs, Artificial neural networks; *CABG,* coronary artery bypass grafting. *Valve procedure with or without CABG surgery.

risk stratify individual patients.[25] Most risk scoring systems have been developed using a biostatistical method based on a generalized linear model. Different methods to improve the accuracy of risk algorithms have been suggested (eg, include more patients with higher risk, select and identify the most important risk factors, and the use of new algorithmic models such as machine learning techniques of which ANNs are an example).[26-28]

Only a few studies have investigated ANNs in the prediction of survival after cardiac surgery.[9-14] Most of these are based on CABG-only patients[9,11-14] and only 1 included all cardiac surgical procedures.[10] None of these studies found any considerable improvement over the traditional biostatistical methods. Orr[10] and Tu and colleagues[9] showed that an ANN could be used to estimate cardiac surgical mortality, but the performance was equivalent to that of logistic regression. These studies were made on smaller cohorts than the present (1477 and 4782 patients) and used few risk variables (7 and 11). Lippmann and Shahian[11] obtained a similar result when ANNs were used on patients from the Society of Thoracic Surgeons database. Despite that 80,000 patients with 32 risk variables were included in the study, the ANNs showed a performance equivalent to the other prediction models. The authors concluded that no complex nonlinear relationship exists, at least not among the presented risk variables. Similar to the other studies, almost all variables were categorical, and the variable selection was performed in a classic way, by significance testing (*P* value). However, identifying a nonlinear relationship is more likely in continuous than categorical variables. Important risk variables for ANNs may also go unrecognized if traditional statistical variable selection is used.

One fundamental and controversial question is the number of variables optimally included in a risk model. In the present study a total of 72 variables (11 continuous) were used. No prior variable selection such as significance testing was used; instead, the ANNs ranked every variable in order of its importance for the mortality prediction. In a second step, the total number of variables was minimized to include only variables with a positive contribution to the outcome prediction. The largest ROC area was achieved when the 34 top-ranked variables were included in the model. If more variables were included, the discriminatory power decreased.

Five of the studies of ANNs in cardiac surgery used ROC analysis to describe the accuracy and the discrimination for the different models[9,11-14] and 1 compared the number of correctly classified patients.[10] Even if comparison of ROC curves in a statistically valid fashion to evaluate models remains controversial, the ROC curve is currently the best-developed statistical tool for describing performance.[22] Importantly, the ROC curve for the ANN model is consistently above the logistic EuroSCORE ROC curve, making direct comparison possible. When applying a statistical model to clinical practice, cutoff values for sensitivity and specificity are valuable. ANNs performed significantly better than the logistic EuroSCORE at sensitivity cutoffs of 25%, 50%, and 75%. At a sensitivity of 75%, the ANNs classified 720 more survivors correctly than the logistic EuroSCORE model did.

The predictive accuracy of different risk scoring systems may be influenced by numerous factors, such as differences in variable definitions, management of incomplete data fields, surgical procedure selection criteria, and geographical differences in patient risk factors. The prevalence of risk factors in patients referred for heart surgery may also change over time.

The advantages of ANNs are that they do not require any a priori assumptions or knowledge of underlying frequency distributions, have the capacity to model complex nonlinear relationships, and are robust and tolerant of missing data and input errors.[11]

Earlier studies on risk analysis in cardiac surgery have mostly been developed and validated on isolated CABG-only patients[2,4] or all cardiac surgery.[17] Recently the Northern New England Cardiovascular Disease Study Group presented a risk model for aortic valve surgery and another for mitral valve surgery.[29] Analyses comparing risk score performance in different surgical procedures have been lacking. In the present study, the ANNs show a similar performance independent of the surgical procedure, unlike the logistic EuroSCORE model. This may be explained not only by a better risk factor selection but also by the capacity of ANNs to recognize complex nonlinear relationships.

Strengths of the present study are that the ANNs were developed on a large multi-institutional database from 8 European countries, that the patient data were quality-checked and validated by 2 independent operators before it was entered into the database,[17] that a large number (42,500) of combinations of parameters included in the ANNs architecture could be evaluated by using high-performance computer clusters, and that an independent blind test was performed on a second, external database. A limitation of the present study is that it was performed on data collected 10 years ago. However, a similar result was obtained in the blind test, where the surgical procedures were performed between 1996 and 2001. Hierarchical generalized linear modeling, which accounts for clustering of observations within providers, may improve the results of the logistic regression.[30] This method may be particularly useful to rate provider performance.

The additive EuroSCORE algorithm[5] can be used at the bedside without a computer, and the logistic EuroSCORE[19] is available on the Internet (http://www.euroscore.org). The ANN model cannot compete with the additive model in simplicity, but it is feasible to make it available on a website.

We thank Per-Anders Wernberg for fruitful discussions and assistance with computer questions.

References

1. Nilsson J, Algotsson L, Hoglund P, Luhrs C, Brandt J. EuroSCORE predicts intensive care unit stay and costs of open heart surgery. *Ann Thorac Surg*. 2004;78:1528-34; discussion 1534-5.
2. Eagle KA, Guyton RA, Davidoff R, Ewy GA, Fonger J, Gardner TJ, et al. ACC/AHA guidelines for coronary artery bypass graft surgery: executive summary and recommendations: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (Committee to revise the 1991 guidelines for coronary artery bypass graft surgery). *Circulation*. 1999;100:1464-80.
3. Shroyer AL, Coombs LP, Peterson ED, Eiken MC, DeLong ER, Chen A, et al. The Society of Thoracic Surgeons: 30-day operative mortality and morbidity risk models. *Ann Thorac Surg*. 2003;75:1856-64; discussion 1864-5.
4. Hannan EL, Kilburn H Jr, Racz M, Shields E, Chassin MR. Improving the outcomes of coronary artery bypass surgery in New York State. *JAMA*. 1994;271:761-6.
5. Nashef SA, Roques F, Michel P, Gauducheau E, Lemeshow A, Salamon R. European system for cardiac operative risk evaluation (EuroSCORE). *Eur J Cardiothorac Surg*. 1999;16:9-13.
6. Burke HB, Goodman PH, Rosen DB, Henson DE, Weinstein JN, Harrell FE Jr, et al. Artificial neural networks improve the accuracy of cancer survival prediction. *Cancer*. 1997;79:857-62.
7. Heden B, Ohlin H, Rittner R, Edenbrandt L. Acute myocardial infarction detected in the 12-lead ECG by artificial neural networks. *Circulation*. 1997;96:1798-802.
8. Baxt WG. Application of artificial neural networks to clinical medicine. *Lancet*. 1995;346:1135-8.
9. 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. *Med Decis Making*. 1998;18:229-35.
10. Orr RK. Use of a probabilistic neural network to estimate the risk of mortality after cardiac surgery. *Med Decis Making*. 1997;17:178-85.
11. Lippmann RP, Shahian DM. Coronary artery bypass risk prediction using neural networks. *Ann Thorac Surg*. 1997;63:1635-43.
12. Ennett CM, Frize M. Weight-elimination neural networks applied to coronary surgery mortality prediction. *IEEE Trans Inf Technol Biomed*. 2003;7:86-92.
13. Buzatu DA, Taylor KK, Peret DC, Darsey JA, Lang NP. The determination of cardiac surgical risk using artificial neural networks. *J Surg Res*. 2001;95:61-6.
14. Chong CF, Li YC, Wang TL, Chang H. Stratification of adverse outcomes by preoperative risk factors in coronary artery bypass graft patients: an artificial neural network prediction model. *Proc AMIA Symp*. 2003:Nov;160-4.
15. Warren S. How to measure importance of inputs? 2000, SAS Institute Inc. June 23, 2000. Available at: ftp://ftp.sas.com/pub/neural/importance.html. Accessed June 23, 2005.
16. Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, Westermann F, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat Med*. 2001;7:673-9.
17. Roques F, Nashef SA, Michel P, Gauducheau E, de Vincentiis C, Baudet E, et al. Risk factors and outcome in European cardiac surgery: analysis of the EuroSCORE multinational database of 19030 patients. *Eur J Cardiothorac Surg*. 1999;15:816-22; discussion 822-3.
18. Cross SS, Harrison RF, Kennedy RL. Introduction to neural networks. *Lancet*. 1995;346:1075-9.
19. Michel P, Roques F, Nashef SA. Logistic or additive EuroSCORE for high-risk patients? *Eur J Cardiothorac Surg*. 2003;23:684-7.
20. Hosmer DW, Lemeshow S. Model-building strategies and methods for logistic regression: In: Hosmer DW, Lemeshow S, editors. Applied logistic regression, 2 ed. New York: John Wiley & Sons; 2000. p. 91-134.
21. Wehrens R, Putter H, Buydens L. The bootstrap: a tutorial. *Chemom Intell Lab Syst*. 2000;54:35-52.
22. Pepe MS. The receiver operating characteristic curve. In: Pepe MS. The statistical evaluation of medical tests for classification and prediction. New York: Oxford University Press; 2003. p. 92-94.
23. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988;44:837-45.
24. 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.
25. Pinna-Pintor P, Bobbio M, Colangelo S, Veglia F, Giammaria M, Cuni D, et al. Inaccuracy of four coronary surgery risk-adjusted models to predict mortality in individual patients. *Eur J Cardiothorac Surg*. 2002;21:199-204.
26. Jones RH, Hannan EL, Hammermeister KE, Delong ER, O'Connor GR, Luepker RV, et al. 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.
27. Omar RZ, Ambler G, Royston P, Eliahoo J, Taylor KM. Cardiac surgery risk modeling for mortality: a review of current practice and suggestions for improvement. *Ann Thorac Surg*. 2004;77:2232-7.
28. Wyse RK, Taylor KM. Using the STS and multinational cardiac surgical databases to establish risk-adjusted benchmarks for clinical outcomes. *Heart Surg Forum*. 2002;5:258-64.
29. Nowicki ER, Birkmeyer NJ, Weintraub RW, Leavitt BJ, Sanders JH, Dacey LJ, et al. Multivariable prediction of in-hospital mortality associated with aortic and mitral valve surgery in Northern New England. *Ann Thorac Surg*. 2004;77:1966-77.
30. Shahian DM, Normand SL, Torchiana DF, Lewis SM, Pastore JO, Kuntz RE, et al. Cardiac surgery report cards: comprehensive review and statistical critique. *Ann Thorac Surg*. 2001;72:2155-68.

Appendix E1

Calibration and Validation of the Artificial Neural Network Models

An ensemble approach was used, where several artificial neural networks (ANNs) were combined into a single prediction model. The individual members of the ensemble were standard multilayer perceptrons (MLPs) with 1 hidden layer and 1 output node that was used to encode the operative mortality.[1] Each MLP was trained using conjugate gradient descent applied to a mean square error function. To avoid overtraining and improve the generalization performance, a weight decay regularization term was utilized. The output of the neural network ensemble was computed as the mean of the output of the individual members in the ensemble. The model selection (ie, the procedure to find the optimal set of model parameters and to select important risk variables) was performed using a sixfold cross-validation procedure (see Figure 1). The model selection procedure explored a large number of different parameter settings by using high-performance computer clusters, including the size of the MLPs, weight decay parameters, size of the ensemble, and risk variable selection (see below).

The final prediction models were tested on patients not previously exposed to the ANNs or the logistic EuroSCORE, by using a fourfold cross-testing technique. Thus, the patient material was randomly split into 4 groups. One of these groups was selected as a test set and excluded from further analysis. The remaining groups were used for calibration and validation. This procedure was performed 4 times with a new group selected each time as a test set (see Figure 1).

Selection of Risk Factors Utilized for Mortality Prediction

To select the most important risk variables and to minimize the number of variables included in the final model, a risk variable ranking was performed. A baseline receiver operating characteristics (ROC) area (see below) was created using all 72 variables. The ranking list was then obtained by measuring the change of the ROC area, as compared with the baseline, when a risk variable was excluded from the model. The highest-ranked variable corresponded to the largest decrease of the ROC area when it was excluded from the model. Each of the models lacking one of the risk variables was recalibrated prior to the ROC area assessment. To optimize the model an increasing number of the ranked variables was included in the model, starting with the top-ranked variable. In this procedure the ANNs were recalibrated after every second variable inclusion.

Effective Odds Ratio and Confidence Intervals for the ANN Output

The odds ratio for a specific risk variable in each patient was determined by changing the risk variable in the patient from "absent" to "present" and calculating the odds for the two conditions. By computing the geometric mean for the odds ratio from all patients, an effective odds ratio for the specific variable was obtained.[2] The 95% confidence limit for both the odds ratio and the output from the ANNs were calculated using the bootstrap technique.[2,3]

From the original database, 1750 bootstrap training data sets were created by resampling with replacement. These bootstrap training sets were then used to calibrate new ANN models with the same architecture and parameters settings as for the final ANN risk prediction model. Each ANN model generated an odds ratio for each variable, resulting in 1750 odds ratios for each variable. Standard techniques[2,3] were then used to extract the confidence limits from these sets of odds ratios. The confidence limits for the mortality risk of individual patients were calculated in the same way.

Computer Cluster

Three clusters for high-performance computing were used to train and evaluate the ANNs. Two Linux clusters hosted by Lunarc at Lund University, one with 210 AMD Opteron nodes and one with 184 Intel P4 nodes, and one Mac OS X cluster with 7 nodes were employed. The latter was also used for the statistical analysis.

References E1

1. Cross SS, Harrison RD, Kennedy RL. Introduction to neural networks. *Lancet*. 1995;346:1075-9.
2. Lippmann RP, Shahian DM. Coronary artery bypass risk prediction using neural networks. *Ann Thorac Surg*. 1997;63:1635-43.
3. Wehrens R, Putter H, Buydens L. The bootstrap: a tutorial. *Chemom Intell Lab Syst*. 2000;54:35-52.