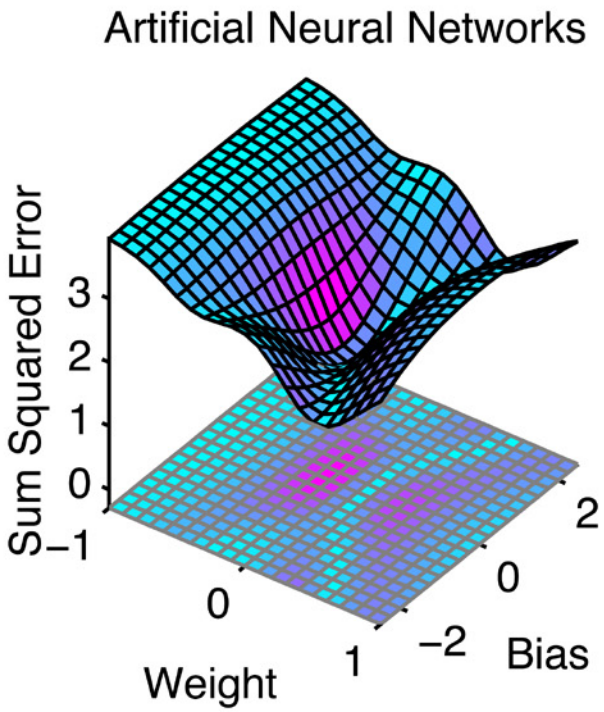
Risk stratification in cardiac surgery: Algorithms and applications

Nilsson, Johan

2005

Risk Stratification in Cardiac Surgery

Algorithms and Applications



Johan Nilsson

Department of Cardiothoracic Surgery
Faculty of Medicine
Lund University
2005

Risk Stratification in Cardiac Surgery

Algorithms and Applications

JOHAN NILSSON, M.D.

Department of Cardiothoracic Surgery
Faculty of Medicine

Lund University
Sweden 2005

Doctoral Dissertation

Department of Cardiothoracic Surgery
Medical Faculty
Lund University
SE-221 85 Lund
SWEDEN

Lund University
Printed by Media-Tryck, Lund, 2005
ISBN 91-85481-03-3

To Bodil,
Sofie and Hanna

"Not everything that counts can be counted, and not everything that can be counted counts."

Albert Einstein

Contents

List of Publications

This thesis is based on the following papers, which are referred to in the text by their Roman numerals:

I. Nilsson J, Algotsson L, Höglund P, Lührs C, Brandt J. Early mortality in coronary bypass surgery: The EuroSCORE versus the Society of Thoracic Surgeons risk algorithm. Ann Thorac Surg 2004;77:1235-9.

II. Nilsson J, Algotsson L, Höglund P, Lührs C, Brandt J. A comparison of nineteen preoperative risk stratification models in open-heart surgery. Submitted.

III. Nilsson J, Algotsson L, Höglund P, Lührs C, Brandt J. EuroSCORE predicts intensive care unit stay and costs of open-heart surgery. Ann Thorac Surg 2004;78:1528-34.

IV. Nilsson J, Ohlsson M, Thulin L, Höglund P, Nashef SAM, Brandt J. Risk factor identification and mortality prediction in cardiac surgery using artificial neural networks. Submitted.

Summary

The aims of this research was to compare different risk score algorithms with regard to their validity to predict 30-day and one-year mortality after open-heart surgery, to evaluate if the preoperative risk stratification model EuroSCORE predicts the different components of resource utilization in cardiac surgery, and to systematically evaluate the accuracy and performance of artificial neural networks (ANNs) to select and rank the most important risk factors for operative mortality in open-heart surgery.

Preoperative evaluation of the surgical risk is an important component in cardiac surgery. Risk stratification can give patients and their relatives insight into the existent risk of complications and mortality, and aid in the selection of cases for surgery versus alternative, non-surgical therapies. It may also predict the need for hospital care resources and improve the quality of care. A few comparative studies of different risk algorithms exist, but the relative performance of the risk scoring systems currently used is unclear, and it still remains difficult to risk-stratify individual patients.

The present work identified four cardiac surgical risk models with a superior performance, with the EuroSCORE algorithm performing best. Though the algorithms were originally designed to predict early mortality, the one-year mortality prediction was also reasonably accurate. The additive EuroSCORE algorithm was also shown to be useful to predict intensive care unit (ICU) cost and an ICU stay more than two days after open-heart surgery. In an attempt to improve the mortality prediction further, a machine-learning technique, ANNs, was used. This identified mortality risk factors in a ranked order and defined a minimal set of risk variables resulting in a superior mortality prediction, compared with previously developed algorithms.

Populärvetenskaplig sammanfattning

(Summary in Swedish)

Inför hjärtkirurgi är bedömning av en patients individuella nytta av och risk vid en hjärtoperation en viktig komponent i utredningen inför operationen. En riskstratifiering innebär att patientens olika riskfaktorer (exempelvis sockersjuka, högt blodtryck, antal tidigare hjärtinfarkter, hjärtfunktion) vägs samman till en sannolikhets-bedömning av risken för en viss händelse (t.ex. att inte överleva ingreppet eller att behöva vistas en längre tid på intensivvårds-avdelningen). Risk kan beräknas på olika sätt. Den vanligaste metoden är att varje riskfaktor ger en riskpoäng och att dessa poäng läggs samman. Den sammanlagda riskpoängen ger en uppfattning om riskens storlek. Mer avancerade tekniker beräknar sannolikheten att en viss händelse skall inträffa efter operationen uttryckt i procent. Ett riskstratifieringssystem kan användas för att bedöma den enskilda patientens risk, men även för att bedöma behovet av resurser och för att värdera kvaliteten på ett visst sjukhus och jämföra olika sjukvårdsenheter.

Trots att olika riskbedömningssystem för hjärtkirurgi har funnits sedan slutet på 1980-talet saknas ännu det perfekta systemet. Det är fortfarande svårt att bedöma den enskilda patientens individuella risk, och bra jämförelser mellan de olika systemens förmåga att göra en korrekt riskbedömning saknas.

Målen med detta arbete var att undersöka och jämföra olika riskpoängsystems förmåga att förutsäga om en patient överlever en månad respektive ett år efter en genomgången hjärtoperation, att undersöka om riskpoängssystem kan användas för att förutsäga sjukhusets resursbehov inför en hjärtoperation, samt att använda ett

s.k. artificiellt neuralt nätverk för att identifiera och rangordna de mest betydelsefulla riskfaktorerna och därigenom förbättra risk-bedömningen för den enskilda patienten.

I delarbete I gjordes en jämförelse mellan de två vanligaste riskbedömningssystemens (STS och EuroSCORE) förmåga att förutsäga vilka patienter som kommer att överleva de första trettio dagarna efter en hjärtoperation. I studien analyserades 4497 kranskärlsoperationer. Resultaten visade att EuroSCORE hade den bästa förmågan att särskilja vilka som skulle överleva.

I delarbete II jämfördes 19 olika riskbedömningssystems förmåga att förutsäga vilka patienter som överlever 30 dagar respektive ett år efter operationen. Fyra olika algoritmer visade sig vara bättre än de övriga: logistisk och additiv EuroSCORE, Cleveland Clinic och Magovern. Trots att algoritmerna är utvecklade för att förutsäga 30 dagars överlevnad fungerade de väl för att bedöma även ett års överlevnad.

I delarbete III undersöktes EuroSCOREs förmåga att förutsäga resursåtgången vid hjärtkirurgi. Data från 3413 hjärtoperationer analyserades. Resultaten visade att EuroSCORE med god precision kunde förutsäga den totala kostnaden samt vårdtiden på intensivvårdsavdelningen.

I delarbete IV utvecklades en algoritm bestående av 144 olika neurala nätverk som identifierade och rangordnade olika riskfaktorer. I denna studie analyserades 18362 patienter som genomgått hjärtkirurgi vid sammanlagt 128 europeiska hjärtkliniker. Resultaten visade på en förbättrad förmåga att förutsäga utgången för den enskilda patienten jämfört med EuroSCORE-algoritmen, samt att denna prediktion var oberoende av vilken typ av hjärtoperation som patienten genomgick.

Sammanfattningsvis är EuroSCORE en riskbedömningsalgoritm som fungerar väl för att förutsäga överlevnaden efter 30 dagar och ett år. EuroSCORE kan även förutsäga resursåtgång vid hjärtkirurgi. Bedömningen av den individuella patientens risk kan förbättras ytterligare genom användningen av artificiella neurala nätverk.

Abbreviations

ACC/AHA	American College of Cardiology/American Heart Association
AIDS	acquired immune deficiency syndrome
ANN	artificial neural network
ASA	acetylsalicylic acid
BSA	body surface area
CABG	coronary artery bypass grafting
CAD	coronary artery disease
CCS	Canadian Cardiovascular Society
CHF	congestive heart failure
CI	confidence interval
COPD	chronic obstructive pulmonary disease
CVA	cerebrovascular accident
Hb	haemoglobin
HIV	human immune deficiency virus
IABP	intra-aortic balloon pump
ICU	intensive care unit
IMA	internal mammary artery
LM	left main coronary artery
LOS	length of stay
LV	left ventricular
ECG	electrocardiogram
EF	ejection fraction
N/A	not available
NCD	National Cardiac Database
NNE	Northern New England
NYHA	New York Heart Association
NYS	New York State

MI	myocardial infarction
MLP	multi-layer perceptron
PACCN	Provincial Adult Cardiac Care Network
PAP	pulmonary arterial pressure
PTCA	percutaneous transluminal coronary angioplasty
PVD	peripheral vessel disease
ROC	receiver operating characteristic
SD	standard deviation
STS	Society of Thoracic Surgeons
TIA	transient ischemic attack
UK	United Kingdom
US	United States of America
VA	Veterans Affairs
WBC	white blood cell count

Chapter 1

Introduction

> "It may seem a strange principle to enunciate as the very first requirement in a Hospital that it should do the sick no harm. It is quite necessary, nevertheless, to lay down such a principle, because the actual mortality in hospitals … is very much higher than any calculation founded on the mortality of the same class of diseases among patients treated out of hospital would lead us to expect."
>
> Florence Nightingale, 1863

1.1 Historical notes

One of the earliest advocates of analysing outcome data was Florence Nightingale (1820-1910). She noted a difference in mortality rate between hospitals, with lower mortality in the smaller county hospitals (39%) compared with the larger hospitals in London (91%)[1]. Hospital mortality rates had been tracked in England since the 1600s, but Florence Nightingale made the important observation that crude mortality is not an accurate reflection of outcome. She suggested that not only patient outcomes but also the severity of the disease should be measured[1].

Ernest Amory Codman (1869-1940), a Boston surgeon, was a pioneer in the search for causes of complications. As a medical student, he became interested in outcome analysis after making a bet with his classmate, Harvey Cushing. They challenged each other who could obtain the best outcomes for their patients, when they

administered anaesthesia at the Massachusetts General Hospital[2]. This led them to create anaesthesia records. Codman linked specific outcomes to defined interventions, and deduced that most unfavourable outcomes were results of errors or omissions by physicians[2].

Both Codman and Nightingale viewed outcome analysis as an intermediate step toward the improvement of patient care. Further definition of outcome assessment occurred in the mid 1900s. As different treatment options emerged, it became important to determine the best alternative among multiple therapies. Controlled randomized trials and tests of therapeutical effectiveness were initiated.

In the 1930s, Archibald Leman Cochrane (1909-1988), used the evidence gained from randomized controlled trials to decide the best treatment. This was the beginning of "evidence-based medicine". In 1972 his influential book "Effectiveness and Efficiency: Random Reflections on Health Services" was published. The principles Cochrane set out were straightforward: Because resources would always be limited, they should be used to provide forms of health care which had been shown in properly designed evaluations to be effective. In particular, he stressed the importance of using evidence from randomized controlled trials, because these were likely to provide much more reliable information than other sources of evidence[3]. The year before Cochrane died, he referred to a systematical review of randomized controlled trials of care during pregnancy and childbirth, and suggested that other specialties should apply the method used. In 1992 the first Cochrane centre was opened, and the Cochrane Collaboration was founded in the following year. The Cochrane website (http://www.cochrane.org) provides summaries of available randomized controlled trials on a wide range of medical subjects.

In 1986 the US Health Care Financing Administration promoted public release of crude mortality statistics for open-heart surgery. Providers correctly argued that such data did not account for differences in disease severity between patients[4], and this led to the development of a number of clinical databases and risk models. Soon after this release of unadjusted outcome data the STS established an

Ad Hoc Committee on risk factors for coronary bypass surgery[5], and the development of the STS NCD was started (http://www.sts.org). The database was established in 1989 and the collection of patient data was started in 1990. Simultaneously other US centres established cardiac databases, such as the NNE Cardiovascular Disease Study Group Cardiac Database[6], the NYS Department of Health[7] and the VA Administration Database[8]. In 1991, the Ontario Ministry of Health established the PACCN, a province-wide computerized registry for monitoring cardiac surgery in Canada[9].

In Europe several cardiac surgical databases were established, such as the UK national database[10] and the Swedish cardiac surgical database (http://www.ucr.uu.se/hjartkirurgi/index.htm). Most of these were based on the STS data format[11].

1.2 Risk stratification

Once a patient is a candidate for cardiac surgery, an important part of the preoperative preparation is assessment of the surgical risk. Risk stratification, defined as the ability to predict outcome from a given intervention by arranging patients according to the severity of their illness, can provide information to patients and their relatives about the chance of undergoing certain procedures successfully, and the risk of complications.

Preoperative risk assessment can be used in patient management (counselling and treatment selection)[12, 13], to improve and compare provider performance (profile provider quality)[8, 14, 15] and for research (e.g. to assess the impact of specific predictors on outcome)[16, 17].

Several risk stratification systems have been developed during the last decades, with the aim to find standardized criteria for comparing outcome in relation to preoperative conditions. An overview of risk models in cardiac surgery is shown in Table 1.1.

The usefulness of any risk stratification system is determined by how well the system connects risk factors to a specific outcome[18]. Although it is almost 20 years since Victor Parsonnet[19] published one of the first risk-score models in open-heart surgery, and despite the

fact that risk prediction models are constantly evolving, the accuracy of risk stratification remains problematic.

Table 1.1. Examples of risk stratification systems used for patients undergoing cardiac surgical procedures. Modified from Cohn and Edmunds: Cardiac surgery in the adult (http://cardiacsurgery.ctsnetbooks.org/)[18].

Scoring system	Data source	Classification approach	Outcomes measured
APACHE III[20]	Values of 17 physiologic parameters and other clinical information	Integer scores from 0 to 299 measured within 24 hours of ICU admission	In-hospital death
NYS[7, 14, 21]	Condition-specific clinical variables from discharge record	Probability of in-hospital death ranging from 0 to 1 based on logistic regression model	In-hospital death
STS[11]	Condition-specific clinical variables from discharge record	Bayesian algorithm used to assign patient to risk interval (percent mortality interval); more recently converted to logistic regression model	In-hospital death and morbidity
VA[8, 22, 23]	Condition-specific clinical variables measured 30 days after operation	Logistic regression model used to assign patient to risk interval (percent mortality interval)	In-hospital death and morbidity
Parsonnet[19]	Condition-specific clinical variables from discharge record	Additive multiple regression model with scores between 0 and 158 based on 16 weighted risk factors	Death within 30 days of operation
Ontario[9]	Condition-specific clinical variables entered at time of referral for cardiac surgery	Range of scores from 0 to 16 based on logistic regression odds ratio for 6 key risk factors	In-hospital mortality, ICU stay, and postoperative length of stay
NNE[6, 24, 25]	Condition-specific clinical variables and comorbidity index entered from discharge record	Scoring system based on logistic regression coefficients used to calculate probability of operative mortality from 7 clinical variables and 1 comorbidity index	In-hospital mortality
Cleveland Clinic[26]	Condition-specific clinical variables from discharge record	Range of scores from 0 to 33 based on univariate odds ratio for each of 13 risk factors	In-hospital death or death within 30 days of operation

One important reason for failure of risk-adjustment methods to completely predict outcomes is that the dataset used to derive the risk score comes from retrospective, observational data, which contain inherent selection bias[18]. Patients are not allocated to treatments in a randomized manner. Rather, treatment recommendations are based on the physician's conviction that a certain therapy is appropriate for the individual patient.

Identification of the best performing risk algorithms is crucial to accurate preoperative risk stratification, and thus also to make fair comparisons. An ultimate goal of risk stratification is to account for differences in patient risk factors so that patient outcome may be used as an indicator of quality of care. However, no universally accepted definition of quality of care exists.

1.3 Outcome analysis

Medical outcome data are often used to compare treatments or providers of care. In clinical surgery there are at least four outcomes of interest: mortality, serious nonfatal morbidity, resource utilization and patient satisfaction. Which of the patient characteristics that represent important risk factors depends on the outcome of interest. For example, mortality risk after open-heart surgery is associated with left ventricular ejection fraction, emergency surgery and recent myocardial infarction[27], whereas increased resource utilization is associated with comorbidity such as peripheral vascular disease, renal dysfunction and chronic pulmonary disease[28-31].

Most risk scoring systems in cardiac surgery have been developed to predict mortality after surgery. Operative mortality is an easily defined and readily measured outcome, and has been widely used as an indicator of the quality of cardiac surgery[32-34]. However, outcomes such as postoperative quality of life or resource utilization may be as relevant, particularly when deciding how to allocate health care resources.

1.4 Databases

In the development of a risk stratification model, a database with patient data is a prerequisite. The importance of the database quality cannot be overemphasized. Factors such as the sources of data, standardized definitions of the data variables, outcomes of interest, data collection methodology, data reliability checking and the time frame of data collection are essential features when using an existing database or constructing a new one[35, 36].

1.4.1 Sources of data

No risk adjustment model is better than the data from which it is developed. Administrative data such as Centres for Medicare data was a commonly used source for observational studies. Such data are readily available, inexpensive and contain information of millions of patients[37, 38]. However, because these data are generated primarily for billing purposes (claims) rather than for clinical studies, their clinical accuracy is inadequate[38]. Critical variables such as left ventricular ejection fraction, emergency surgery and recent myocardial infarction are often missing. Such administrative data have been found to underestimate the effect of comorbid illness. The Duke Databank for Cardiovascular Disease found major discrepancies between clinical and administrative databases, with claims data failing to identify more than half of the patients with important comorbid conditions[39]. Today, clinical databases are therefore generally used for risk stratification and outcome analysis. An example is the STS NCD, which includes over 2 million patients[11, 40].

1.4.2 Variable definition

Strict standardization of definitions, both for predictor variables and endpoint variables, is essential in a successful risk stratification model. Even for an apparently obvious endpoint (such as mortality) there are important considerations. Mortality could be defined as in-hospital mortality (regardless of when it occurs), 30-day mortality (regardless of where it occurs) and operative mortality (including both in-hospital and 30-day mortality). A fixed time period is preferable from a statistical point of view[35], but may be more difficult to use than in-hospital mortality.

Osswald and coworkers[33] have studied the implications of different definitions of early postoperative mortality, especially in the light of improvements in postoperative care. For high-risk patients, with a prolonged early postoperative phase, mortality resulting from surgery may be underestimated. A definition of operative mortality as occurring within 6 or 12 months may be more relevant. For other endpoints, such as postoperative morbidity, postoperative complications and hospital cost standardized definitions are even more difficult.

1.4.3 Quality

In addition to variable definition, the type of data collected is important. Continuous data should be used to avoid arbitrariness and loss of valuable information that may occur with categorization[41]. If this is impractical or a categorization may be useful to identify cases (e.g. patients with renal insufficiency or morbid obesity) it is important that the categorical state is well defined and widely accepted.

The accuracy of the risk stratification model is dependent on the condition of the input data, and the quality of the data entry is important. The data entry software should contain internal quality controls for out-of-range, inconsistent and missing data. In situations where risk-adjusted outcomes are used to assess provider performance, there should be regular and independent auditing of the data to assure accuracy and completeness. The problem of missing data should be avoided as far as possible. If this situation still occurs, missing data values may be substituted using imputation techniques[42, 43].

1.5 Strategies for model development

Strategies for model development[44-48] include:

- Variable transformation
- Imputation of missing values
- Variable selection
- Model techniques
- Validation of risk stratification models

1.5.1 Variable transformation

Variable transformation denotes a change in the scale of the measurement of a variable during model development. Common reasons for variable transformation are variance stabilization, linearization, normalization, simplification of data handling, and to enable more appropriate presentation of the results[48].

1.5.2 Imputation of missing values

The analysis of prognostic factor studies is often impeded by missing covariate values. The most common strategies is still to omit incomplete cases and/or to delete covariates from the analysis, both with known undesirable effects[49]. In a review, Little[43] concluded that among methods available, model-based approaches, such as maximum likelihood estimation, Bayesian methods[50] and multiple imputation[42] are preferable. Unfortunately there are no corresponding easily applied software tools available for these techniques. Schemper and Smith[49, 51] have suggested a simpler probability imputation technique, which substitutes conditional probabilities for missing covariate values when the covariate is qualitative.

1.5.3 Variable selection

The optimal number of variables to include in a risk model is a controversial question. Too many variables may to lead over-fitting of the model[46] (i.e. an extremely good fit to the risk model database but limited ability to predict future events), instability, increased cost and difficulty of data collection. Too few variables may decrease the performance of the model (under-fitting). Harrell and coworkers[47]

have recommended that the number of risk variables considered for inclusion should be less than one-tenth the number of cases with the defined outcome in the model data set.

The most common data reduction technique is univariate screening (significance test[52]) of risk variables such as Student's t-test or χ^2 test, followed by forward or backward selection (e.g. stepwise backward logistic regression analysis[46]). These techniques are generally available in commercial statistical software. Other techniques are principal component analysis[53], bootstrap bagging[54, 55] and risk variable ranking[52].

A risk variable ranking[52] is performed by calculating a baseline performance (e.g. mean square error or area under the ROC curve) using all available risk variables. The ranking list is then obtained by measuring the change of the performance, as compared to the baseline performance, when a risk variable is excluded from the model. The highest ranked variable corresponds to the largest decrease in performance when it is excluded from the model. To optimize the model an increasing number of the ranked variables are included in the model, starting with the top ranked variable. The algorithm is recalibrated after each variable inclusion and the performance is calculated.

1.5.4 Model techniques

Several techniques are available for risk model development. All require considerable statistical knowledge, which is sometimes forgotten in this era of powerful off-the shelf statistical software. Three principal techniques have been utilized for construction of cardiac surgery risk models: Bayesian analysis, regression analysis and machine-learning techniques.

Bayesian analysis

Bayesian analysis is a statistical procedure to estimate parameters of an underlying distribution based on the observed distribution. Bayes' theorem[50] provided a mathematical method that could be used to calculate, given occurrences in prior trials, the likelihood of a target occurrence in future trials (Table 1.2). Bayes' theorem is a means of quantifying uncertainty. The principle of Bayesian technique has been used widely in decision analysis[56] and can be used to generate

multivariable regression models. It was used to develop the original risk stratification analysis for the STS NCD[11]. The Bayesian technique is robust regarding missing values, which was an important problem in the early database experience. Marshall and coworkers[56] have shown that the Bayesian model of risk adjustment gives results comparable to those generated from logistic regression analysis. Since 1995, the STS NCD uses the logistic regression model[57].

Table 1.2. Equations used in different model development techniques.

Logistic regression analysis	$P = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 .... \beta_n x_n}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 .... \beta_n x_n}}$
Bayesian analysis	$P(A \mid B) = \frac{P(B \mid A) P(A)}{P(B)}$
Artificial neural network	
Output function	$Y_k = \varphi_o \sum_j^H \omega_{kj} \varphi_h (\sum_i^I \tilde{\omega}_{ji} x_i)$
Error function	$E = \frac{1}{2N} \sum_{n=1}^{N} \sum_{i=1}^{C} (d_i(n) - y_i(n))^2$
Update function	$\Delta\omega = -\eta \frac{\partial E}{\partial \omega}$
Generalisation function	$\tilde{E} = E + \alpha \frac{1}{2} \sum_i \omega_i^2$

Regression analysis technique

Regression analysis is the most commonly used statistical technique for cardiac surgical risk estimation[58]. The analysis describes how one variable depends on or is associated with a set of independent (or predictor) variables. The outcome can be a continuous variable or dichotomous. The significant, independent variables are termed risk factors. Knowledge of these risk factors allows separation of patients according to their degree of risk – i.e. risk stratification. For a continuous outcome (e.g. length of stay) a linear regression analysis is

used, and for dichotomous (e.g. mortality) a logistic regression analysis is applied[48].

Logistic regression analysis uses the past experience of a patient group to estimate the odds of an outcome by mathematically modelling or simulating that experience. The method of calculation for the regression coefficients takes into consideration all possible combinations of the independent variables. It maximizes the probability that, for any given individual with a particular combination of independent variables, the odds of the outcome will be close to the actual or observed outcome of all other individuals with the same combination of independent variables[46]. The general form of the logistic regression equation is similar to that of multivariable linear regression, with the exception that the logarithm of the odds of the outcome is used as the dependent variable (Table 1.2). The exponential of the regression coefficients for each predictor variables are the odds that the outcome will occur if the predictor variable is present, compared with if it is absent. For example, the odds ratio for the variable "prior cardiac operation" is 2.76 in the logistic EuroSCORE[59] model. If the operative mortality in the absence of risk factors is 0.6%, then the operative mortality for a patient with a prior cardiac operation (but no other risk factors) is 1.7% (0.6 x 2.76). If the patient also has a "neurological dysfunction" with the odds ratio of 2.3, the mortality risk will be 3.8% (0.6 x 2.76 x 2.3). In the construction of an additive risk-score model the additive scores are weight-derived from the logistic model, normally by rounding off the odds for the predictors to nearest integer. The risk is predicted by adding the score for each risk factor. In the above example, the mortality risk in the additive EuroSCORE[60] model for a patient with "prior cardiac surgery" (3 points) and "neurological dysfunction" (2 points) will be 5% (3 + 2).

Traditional logistic regression modelling to rank surgeons according to their risk-adjusted mortality rates may result in incorrect provider profiles[61-63]. In order to grade a treatment provider, the expected number of deaths (calculated from deaths observed in the entire provider group) is compared with the observed number of risk-adjusted deaths for the provider. This gives a ratio of the risk-adjusted observed mortality rate to the expected mortality rate, based on the group logistic model.

The expected mortality rate is assumed to be independent of the observed mortality rate and no sampling error is attached to the expected values; however, these two assumptions are incorrect[18, 63]. The effect is that too many providers are identified as outliers[63, 64]. The methodology to account for this problem involves construction of hierarchical logistic regression models[63, 65], that incorporate (nest) other levels of analysis within the analysis. For example, patients are nested in provider groups (cases treated by a given provider), but patients are also nested within hospital groups (cases treated at a given hospital). A hierarchical model[63, 65] recognizes that the nested variables may be correlated (e.g., mortality may depend on the surgeon, the hospital where care is provided, and other unspecified variables such as referral patterns, hospital size[66] and location[67]) and that different sources of variation can occur at each level (or nest). Presently, hierarchical regression is the "gold standard" for risk adjustment of dichotomous outcomes. Unfortunately it is rarely used, presumably because the method is complex and not included in most commercially available software[18].

Machine-learning techniques

One of the most promising newer risk-adjustment methods is the use of artificial neural networks. An ANN[68] consists of a set of processing units (nodes) that simulate neurons and are interconnected via a set of weights (analogous to synaptic connections in the nervous system) in a way that allows signals to travel through the network in parallel as well as serially. The nodes are very simple computing elements and are based on the observation that a neuron behaves like a switch: when sufficient neurotransmitter has accumulated in the cell body, an action potential is generated. This has been modelled mathematically (Table 1.2) as a weighted sum of all incoming signals to a node. The weighted sum of the signals is compared with a threshold. If the threshold is exceeded the node fires, otherwise it remains inactive. Computational power in a neural network does not derive from the complexity of each processing unit, but from the density and complexity of the interconnections. The feed forward neural networks, MLP (Figure 1.1) is a popular and widely used neural network model[53]. It uses one or more hidden layers of nodes with an activation function. The learning is usually achieved by minimizing an

error function of the input and target data (Table 1.2). The simplest method for this is "gradient descent". Other methods are "conjugate gradient" and second order methods such as "Levenberg-Marquardt". The best network architecture for a particular task must be developed by experimentation and observation.

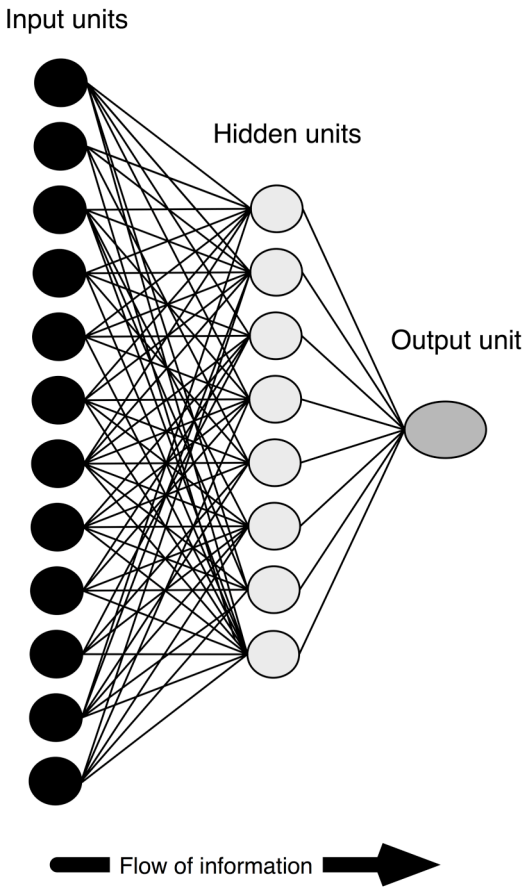


Figure 1.1. Schematic diagram of a multilayer perceptron ANN.

A weakness of multivariable regression techniques is that some variables are present in too low frequency to be used in a multivariable regression model, but still contribute significantly to the outcome[69, 70]. ANNs use computer iteration to look for patterns in different variables associated with outcome and are far less affected by low frequencies in the variables[69, 70]. Further, ANNs work in a nonlinear fashion (Figure 1.2), which may better describe the interaction between health risk factors. ANNs have found medical applications, such as cancer survival prediction[71], screening of heart murmurs in children[72], ECG interpretation[73], etc. Some studies in clinical medicine have demonstrated superiority of the prediction by ANNs compared with other statistical methods[74]. In the field of cardiac surgery, only a few studies using ANNs have been published, and the results have been ambiguous[69, 70, 75-78].

Traditional neural network approaches have suffered difficulties with generalization, producing models that can over-fit the data. The Support Vector Machine is gaining popularity due to many attractive

features, such as better ability to generalize and promising empirical performance[79].

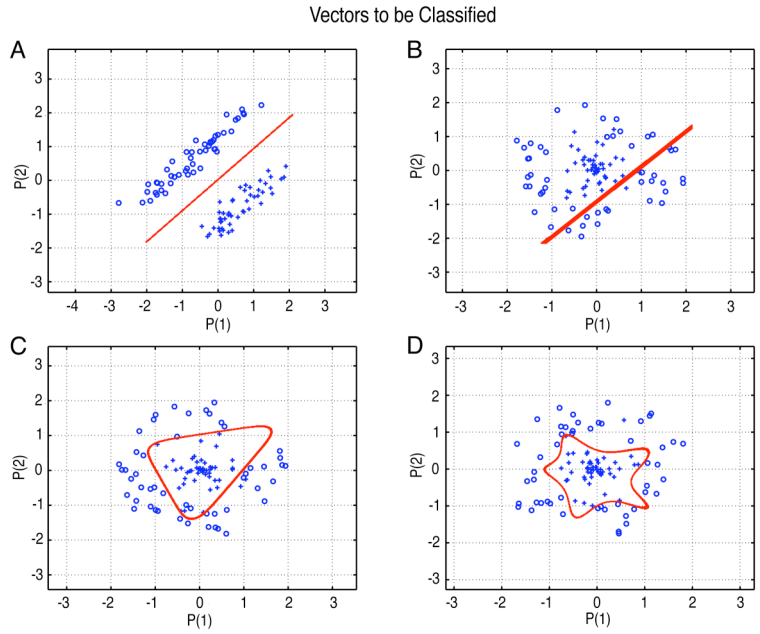


Figure 1.2. Solution of a two-dimensional classification problem. (A) Complete linear separation of the two classes, using a bio-statistical method based on a generalized linear model with assumptions of linear relationship, a rare or non-existent situation in medicine. (B) Poor linear separation of a non-linear classification problem, performed by the same linear model. (C) An almost complete non-linear separation of the two classes, using a bio-statistical method based on a non-linear model such as an ANN with four hidden nodes. (D) Complete non-linear separation of the two classes, using an ANN with eight hidden nodes.

1.5.5 Validation of risk stratification models

Any statistical model must be validated to determine whether it performs adequately. Daley[35] have summarized the following aspects of validity:

- Face validity – Will whoever uses the risk model accept it as valid?
- Content validity - Does the model include risk factors that should have been included, based on known risks?
- Construct validity – How well does the model compare to other measures of the same outcome?
- Attributional validity – Does the model measure the attribute of effectiveness of care, not patient variability?
- Predictive validity – How well does the model predict outcome in patients not used to construct the model?

The predictive validity of a model is a measure of how well it performs on a data set other than the one from which it was developed. This test may be internal or external. An internal test is used on the original data set, whereas an external is made on a separate dataset. Numerous techniques are available to make an internal validation. The original data may be randomly split into a development set and a validation set. Alternatively, more sophisticated techniques may be used: K-fold cross validation[68], leave-one-out cross validation[68] and bootstrapping[54].

K-fold Cross Validation

The final prediction model is tested on patients not previously exposed to the model, by using a K-fold cross validation technique. Thus, the patient material is randomly divided into K groups of equal or roughly equal size. One of these groups is selected as the validation group and excluded from further analysis. The remaining groups are used for model development. This procedure is performed K times with a new group selected each time for validation.

Leave-One-Out Cross Validation

For a dataset with *n* observations, training is conducted on *n* different subsets of data, each of which has one data point left out. Each

excluded observation is predicted by the model obtained from the remaining data points, and the average predictive accuracy of the model is determined[55, 80].

Bootstrapping

From the original database, bootstrap training data sets are created by re-sampling with replacement. A model is developed for each bootstrap sample and tested against those subjects not present in the sample. All bootstrap samples will have the same number of observations as the original, but in each bootstrap sample any of the original observation may be included once, more than once, or not at all. The error is the average over all bootstrap samples. Most authors recommend using at least 1000 bootstrap samples. Bootstrapping can also be used to estimate confidence intervals, and for risk factor identification[54, 58].

1.6 Performance measurement

The accuracy of a model may be evaluated in several ways. The first property is termed "variability" and is a measure of the performance of the risk-adjusted model. The second property is termed "calibration" and is defined as the ability of the model to assign an appropriate risk to the patients upon whom the model is based. A third property relating to the accuracy of a model is termed "discrimination" and is defined as the ability of the model to distinguish between those patients having and those not having the outcome of interest. The fourth property, "reliability", refers to the statistical term precision, i.e. the ability to repeat the observations using similar input variables and similar statistical techniques, with resultant similar outcome findings.

1.6.1 Variability

The strength of the relationship between the dependent and independent variables in a linear regression analysis is given by the correlation coefficient, usually designated as r. The degree of variance explained by the model can be calculated as r^2, which is routinely used as a measure of the performance of linear regression risk-adjusted

models[81]. For a dichotomous output the variance can be calculated as pseudo r^2. For example, an ICU LOS prediction model with a r^2 value of 0.25 implies that 25% of the variability of the ICU LOS can be explained by the model, and that 75% of the variability is not explained by the model. Shwartz and Ash[82] have provided an excellent review of evaluating the performance of risk-adjustment methods, using r^2 as a measure of model performance.

1.6.2 Calibration

Calibration of a model may be assessed by comparing the observed and the expected mortality for equal-sized quantiles of risk groups by using the Hosmer-Lemeshow goodness-of-fit test[46]. This test presents a modified χ^2 statistic, where a non-significant p value for the difference between observed and estimated outcomes is desired. However, models can be adjusted during their development to overcome poor calibration, and many studies will not refer to calibration measures[81].

1.6.3 Discrimination

The potential of the ROC curve in medical diagnostic testing was recognized as early as 1960[83]. The ROC curve is defined as a plot of test sensitivity (as the y coordinate) versus its 1-specificity or false positive rate (as the x coordinate)[84]. The area under the curve is a combined measure of sensitivity and specificity. It is a measure of the overall performance of a model, and it is interpreted as the average value of sensitivity for all possible values of specificity. An area of 1.0 under the ROC curve indicates perfect discrimination, whereas an area of 0.50 indicates complete absence of discrimination. Any intermediate value is a quantitative measure of the ability of the risk predictor model to distinguish between a positive or negative outcome[85]. Sensitivity and specificity are independent of the number of cases with a specific outcome; consequently, so is the ROC analysis[86].

1.6.4 Reliability

The reliability of a risk adjustment method refers to the statistical term precision, or the ability to repeat the observations using similar input variables and similar statistical techniques with resultant similar

outcome findings. The most common measure of reliability is Cohen's *kappa* coefficient, which measures the level of agreement between two or more observations compared to agreement due to chance alone[87]. A *kappa* value greater than 0.75 is often regarded as representing excellent agreement, and 0.4-0.75 as fair to good agreement[88].

1.7 Risk score systems

Since Victor Parsonnet presented "A method of uniform stratification of risk for evaluating the result of surgery in acquired adult heart disease" (the Parsonnet score[19], 1989) numerous risk score models have been developed. The purpose of most of these studies has been to construct a simple, additive risk-score model to predict operative mortality. Frequently cited risk algorithms are the Parsonnet[19], the Cleveland Clinic[26] (also termed Higgins), the Ontario[9] (also termed PACCN) and the additive EuroSCORE[60] risk score model, but there are several more. The risk-score systems presented in Table 1.3 include between 6 and 33 predictors or risk variables. Most models have been developed by the logistic regression technique. In the additive models, each risk variable score is derived from the corresponding odds in the logistic regression analysis (or rounded to the nearest integer). The NNE study group has also recently presented tables where the total risk points can be transformed to a mortality risk proportion[25] (Tables 1.4 and 1.5).

The large cardiac databases in the US use logistic regression analysis to obtain new coefficients for their algorithms yearly[21, 23]. The regression coefficients can be used to calculate risk-adjusted mortality data for the institutions included. The NYS, which publishes risk-adjusted mortality from all of the participating institutions, also publishes the coefficients in the logistic regression algorithm, making it possible for centres to compare their results with others. On the contrary, the proprietors of the largest cardiac surgery database, the STS NCD, have chosen not to publish their algorithm.

Most of the risk score systems have been evaluated and validated on internal data and only a few comparative studies of different risk algorithms have been made[10, 89-91]. Thus, the relative performance of the risk scoring systems currently used has been unclear.

The most common endpoint of cardiac surgical risk scoring systems is the operative mortality, but some investigators have focused on other endpoints such as LOS at the ICU and postoperative morbidity (Table 1.1). A Swedish research group has developed a simple score to assess mortality risk in patients waiting for coronary artery bypass grafting, aiming to improve the prioritization process[92].

Table 1.3. Synopsis of nineteen risk score algorithms.

	Region	Year of data collection	Year of publication	Number of patients (centres)	Risk variables
Amphiascore[93]	Netherlands	1997-2001	2003	7282 (1)	8
Cabdeal[94]	Finland	1990-1991	1996	386 (1)	7
Cleveland Clinic[26]	US	1986-1988	1992	5051 (1)	13
EuroSCORE (additive)[60]	Europe	1995	1999	13302 (128)	17
EuroSCORE (logistic)[59,95]	Europe	1995	2003	13302 (128)	17
French score[96]	France	1993	1995	7181 (42)	13
Magovern[97]	US	1991-1992	1996	1567 (1)	18
NYS[7, 14, 21]	US	1998	2001	18814 (33)	14
NNE[6, 24, 25]	US	1996-1998	1999	7290 (N/A)	8
Ontario[9]	Canada	1991-1993	1995	6213 (9)	6
Parsonnet[19]	US	1982-1987	1989	3500 (1)	16
Parsonnet (modified)[98]	France	1992-1993	1997	6649 (42)	33
Pons[99]	Spain	1994	1997	1309 (7)	11
Toronto[100]	Canada	1993-1996	1999	7491 (2)	9
Toronto (modified)[101]	Canada	1996-1997	2000	1904 (1)	9
Tremblay[102]	Canada	1989-1990	1993	2029 (1)	8
Tuman[103]	US	N/A	1992	3156 (1)	10
UK national score[10]	UK	1995-1996	1998	1774 (2)	19
VA[8, 22, 23]	US	1987-1990	1993	12712 (43)	10

Table 1.4. NNE preoperative estimation of mortality risk in CABG, mitral or aortic valve surgery (NNE Cardio Vascular Disease Study Group, 2004)[104, 105].

Patient or disease characteristics	Mortality score		
	CABG	Aortic valve	Mitral valve
Age 60-69	1.5	1.5	1.5
Age 70-75	2.5	1.5	2.5
Age 76-79	2.5	2	2.5
Age ≥80	6.5	2.5	2.5
Female sex	2	1.5	
EF <40%	2		
NYHA class IV		1.5	2
3-vessel disease	1.5		
LM 50-89%	1.5		
LM >90%	2		
WBC >12000	2.5		
MI <7 days	1.5		
Urgent surgery	2	1.5	1.5
Emergency surgery	5	5	5.5
Prior CVA			2
Prior CABG	2.5	1.5	
CVA, TIA, PVD	1.5		
CHF		1.5	1.5
Atrial fibrillation		1.5	
CAD			1.5
Diabetes	1		1.5
Dialysis	4		
Creatinine >1.3 mg/dL		2	1.5
Creatinine >2.0 mg/dL	2		
COPD	2		
BSA <1.70 m^2		1.5	
Concomitant CABG		1.5	
Mitral valve replacement			1.5

Table 1.5. NNE operative score mortality (NNE Cardio Vascular Disease Study Group 2004)[104, 105].

Total Score	Preoperative risk		
	CABG (%)	Aortic (%)	Mitral (%)
0	0.2		
1	0.2	1	<1.0
2	0.3	1.5	1
3	0.3	2	1.5
4	0.5	3	2
5	0.7	4	2.5
6	1	6	3
7	1.3	7	5
8	1.8	9	6
9	2.3	13	8
10	3	17	11
11	4	20	14
12	5.3	25	18
13	6.9	>35.0	25
14	8.8		>35.0
15	11.5		
16	14.1		
17	18.7		
18	>23.0		

1.8 Limitations and ethical considerations of risk stratification

The causes of adverse outcomes are generally multifactorial. Risk stratification in cardiac surgery has been based mainly on patient-related risk factors. However, factors such as the skill, equipment and organization of health care providers and chance certainly play a role. Risk stratification models based only upon patient characteristics are therefore unlikely to reach absolute predictive accuracy.

Most available risk score systems have been focused on operative mortality prediction. One drawback of present risk models is that they predict the outcome for individual patients poorly[91]. As previously described, the best available risk stratification models explain only a limited proportion of the variability in cardiac surgery. Risk-adjusted mortality data based on these algorithms may be misleading, and may not accurately reflect quality of care. Hopefully, improvements will be seen as new and more sophisticated techniques to model complex biological phenomena are developed.

If risk-adjusted mortality data is used to compare providers of care, this could result in decreased access to surgery for those who might benefit most (high-risk case avoidance)[106, 107] and may encourage "gaming". Gaming may occur in several forms such as upcoding of preoperative comorbidities, unwarranted change of operation procedures, and postoperative transfer of critically ill patients to another unit[63]. High-risk case avoidance may already have occurred as a result of releasing risk-adjusted mortality and cost to the public[108, 109].

Publications of risk-adjusted mortality data in Europe are sparse. A number of institutions (mostly in the UK) have published their operative mortality data. St George's hospital in London has presented the results of individual surgeons, compared with a predicted mortality calculated by the EuroSCORE algorithm on the Internet (http://www.st-georges.org.uk), and so has the Cardiothoracic Centre in Liverpool (http://www.ctc.nhs.uk). The Swedish Association for Thoracic Surgery has chosen to publish the unadjusted mortality data for the Swedish cardiac surgical centres (http://www.ucr.uu.se/hjartkirurgi/index.htm), but not the results of

individual surgeons. Other cardiac centres in Europe will presumably follow.

To prevent providers from selecting only low-risk patients, health plans should allocate resources to health care providers based on the overall patient risk and the associated expected need for resources. The use of risk stratification in this context is new and offers great promise.

Chapter 2

Aims of the Thesis

The general aim of this thesis was to bring risk stratification research and knowledge a step further, in order to achieve a higher quality of treatment and improve the outcome for cardiac surgical patients.

The specific aim for each paper was to

I. compare two widely used risk algorithms for CABG: The EuroSCORE and the STS risk stratification algorithm;

II. compare 19 open-source risk score algorithms with regard to their validity to predict 30-day and one-year mortality after open-heart surgery;

III. evaluate if the preoperative risk stratification model EuroSCORE predicts the different components of resource utilization in cardiac surgery;

IV. 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.

Chapter 3

Material and Methods

The studies were approved by the Ethics Committee of the Medical Faculty, Lund University.

3.1 Databases

Database I

Database I included all adult patients (n=8342) undergoing heart surgery at the Department of Cardiothoracic Surgery, Lund University Hospital between January 1, 1996 and December 31, 2002, and at Malmö University Hospital between January 1, 1996 and December 31, 1997.

Risk factors for all adult patients were prospectively collected when the patients were admitted to the department. The patient record form contained a total of 248 variables (pre-, intra- and postoperative) based on the Higgins[26], Parsonnet[19] and STS[11] patient record forms. The date and cause of mortality was obtained from the Population and Welfare Statistics Sweden (Statistiska Centralbyrån), Stockholm, Sweden.

Database II

Database II (the EuroSCORE database) included 97 risk factors from all adult patients (n=19030) undergoing heart surgery in 128 centres from eight European countries during September to December, 1995. The database was originally used in the multinational 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[110] and to construct the additive[60] and logistic[59, 95] EuroSCORE risk stratification systems. The data collection, quality checks and validation have been described by Roques et al[110].

3.2 Study design

Study I

Twenty-six risk variables included in the STS[11] and EuroSCORE algorithms[60] (Table 3.3) were imported from *Database I* into the commercially available STS risk stratification software (see chapter 4.7) together with the 30-day mortality for the CABG-only population operated between January 1, 1996 and February 28, 2001. The risk stratification software calculated the risk score for every patient according to the STS Risk Stratification Analysis version 2.0 algorithm, which is based on the STS NCD 1990-1993 CABG-only population, and the EuroSCORE additive algorithm[60]. The accuracy and performance of each algorithm were evaluated using the Hosmer-Lemeshow test and ROC-analysis (see chapter 4).

Study II

One-hundred-four of the pre- and intra-operative variables from *Database I* were imported into the statistical software package (see chapter 4.7), together with 30-day and one-year mortality for patients undergoing open-heart surgery between January 1, 1996 and February 28, 2001. Patient characteristics are summarized in Table 3.4a-d. Missing values were replaced using the probability imputation technique[49] before the risk score was calculated. The probability imputation technique substitutes conditional probabilities for missing covariate values when the covariate is qualitative. The risk score for each of the 19 risk score algorithms was calculated for every patient according to the published definitions (Table 1.3). The accuracy and performance of each algorithm were evaluated using ROC-analysis (see chapter 4).

Study III

The 18 EuroSCORE risk variables (Table 3.5) together with duration of anaesthesia (minutes), the Lund ICU workload score, and the LOS

at the ICU, in the ward, and the total in-hospital stay were imported from *Database I* into the statistical software package (see chapter 4.7) for patients undergoing open-heart surgery between October 1, 1999 and December 31, 2002. Patients, who had a heart transplant, died intra-operatively, or in whom any of the pre-, intra- or postoperative data were missing were excluded from the study.

The Lund ICU workload score is a modification of a nursing care recording system[111], by which each patient in the ICU gets a score three times a day, depending on the resources needed for his/her condition (e.g. medication, volume therapy, transfusions, need of ventilator assistance, need of further technical support such as dialysis or cardiac assist device, and nursing workload). Scoring points are directly related to the cost of the specific resource used. The total number of points is computed daily for each patient and entered into the database.

The total risk score for every patient was calculated according to the EuroSCORE additive algorithm[60], and the individual cost according to a formula used by the hospital accounting system. Principles for calculations of costs of care are shown in Table 3.1. The hospital economy department established all starting costs and constants yearly.

The accuracy and performance of the EuroSCORE algorithm was evaluated using the Pearson correlation test, the Hosmer-Lemeshow test and ROC-analysis (see chapter 4). Analyses were performed using both individual patient data and patients grouped into six risk cohorts[29, 30] (Table 3.2).

Table 3.1. Principles for calculation of costs of care.

Costs	Equation
Surgery	Starting $Cost_{Op}$ + (Anesthesia duration x $Constant_{Op}$) + Implantable material
ICU	Starting $Cost_{ICU}$ + (Total Lund ICU workload score x $Constant_{ICU}$)
Ward	Starting $Cost_{Ward}$ + (LOS ward x $Constant_{Ward}$)

Table 3.2. Patient cohorts based on EuroSCORE risk stratification.

Cohort	EuroSCORE risk	No of operations
I	0-2	612
II	3-4	672
III	5-6	700
IV	7-8	614
V	9-10	423
VI	>10	392

Study IV

A subset of 72 variables from the 97 variables included in *Database II* was selected (Table 3.6a-b), by excluding variables closely linked to other variables, and data collected intra-operatively (i.e. 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 also excluded from analysis. Missing values in the other variables were substituted with their most likely values (the statistical mode for categorical variables, and the mean value for continuous variables)[70].

A subset from *Database I,* including risk factors for adult patients undergoing heart surgery between January 1, 1996 and February 28, 2001 was used to further evaluate the developed ANN risk model by external blind testing.

A ranking of risk variables was performed to identify the most important risk factors[52]. The software used and the training and validation procedure for the ANN model is described in chapter 4.4-4.7. The accuracy and performance of the different algorithms were evaluated using ROC-analysis and a proportion test[112].

3.3 Patient characteristics

Study I

Table 3.3. Patient characteristics in 4497 CABG-only operations.

Risk Factors		Prevalence (mean±SD, or %)	EuroSCORE	STS
Age (years)		66.4±9.3	√	√
Female gender		23.0%	√	√
Morbid obesity		12.1%		√
Chronic pulmonary disease		6.9%	√	√
Diabetes		17.8%		√
Extracardiac arteriopathy		11.2%	√	
Cardiomegaly		3.1%		√
Renal failure:	Serum Creatinine >167 µmol	2.6%		√
	Serum Creatinine >200 µmol	1.7%	√	
Neurological dysfunction		7.1%	√	√
Preoperative myocardial infarction:	No MI	36.6%		√
	>21 days	47.0%		√
	<21 days	16.4%	√	
	8-21 days	10.2%		√
	1-7 days	4.8%		√
	6-24 hrs	0.7%		√
	<6 hrs	0.7%		√
Stable angina		79.0%		√
Unstable angina		15.8%	√	√
LV dysfunction:	EF (%)	51±11		√
	EF >50%	63.1%		
	EF 30-50%	30.4%	√	
	EF <30%	6.5%	√	
Pulmonary hypertension		0.8%	√	
Left main disease		22.4%		√
Number of vessels diseased:	One vessel	5.8%		√
	Two vessels	27.6%		√
	Three vessels	66.6%		√
Aortic valve disease		6.5%		√
Mitral valve disease		13.7%		√
Operative incidence:	First operation	94.8%		√
	Redo	5.2%	√	√
PTCA emergency <6 hours		0.4%		√
PTCA emergency >6 hours		0.9%		√
Cardiogenic shock		0.7%		√
Critical state (preoperative)#		3.9%	√	
Elective		66.7%		√
Urgent		25.1%		√
Emergent		7.2%	√	√
Emergent/Salvage		1.0%		√

The check mark indicates the risk variable included in each risk algorithm.
For definition see Table 3.4c.

Study II

Table 3.4a. Preoperative general risk factors in 6222 open-heart operations.

Preoperative Risk Factors	Mean (SD) or n (%)	Amphiascore	Cabdeal	Cleveland Clinic	EuroSCORE*	French score	Magovern	New York State	Northern New England	Ontario	Parsonnet	Parsonnet (modified)	Pons	Toronto	Toronto (modified)	Tremblay	Tuman	UK national score	Veterans Affairs
Age† (years)	66.3 (10.6)	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
Female gender	1765 (28.4)	√			√		√	√	√	√	√	√		√	√		√	√	
Height† (centimetres)	171 (8)		√				√									√			
Weight† (kilograms)	79 (14)		√	√			√				√	√				√			
Hb† (g/L)	134 (16)			√			√												
Serum Creatinine† (μmol/L)	95 (40)	√	√	√	√	√	√	√				√	√				√		
Hypertension (sys >140 mmHg)	2458 (40.0)										√	√		√	√			√	
Diabetes	1106 (17.9)		√	√			√	√			√	√			√			√	
Hypercholesterol-emia (treated)	2274 (37.0)											√							
Chronic pulmonary disease	477 (7.7)		√	√	√		√	√	√			√						√	
Active smoker	539 (8.8)																	√	
Cerebrovascular disease	448 (7.2)			√	√		√	√				√					√	√	√
Peripheral vascular disease	636 (10.3)			√	√		√	√	√			√		√	√			√	√
Kidney disease by history	248 (4.0)																√	√	
Dialysis	28 (0.5)					√		√	√		√	√						√	
Adult congenital heart disease	11 (0.2)											√							
ASA medication	4346 (69.9)											√							
Diuretic medication	2203 (35.4)																		√
Immunosuppressive medication	71 (1.2)											√							

The check mark indicates the risk variable included in each risk algorithm. *Additive and logistic. †Continuous variables are presented as mean (SD). The analysis is based on operations where the risk factor data was available.

Table 3.4b. Preoperative cardiac risk factors in 6222 open-heart operations.

Preoperative Risk Factors	Mean (SD) or n (%)	Amphiascore	Cabdeal	Cleveland Clinic	EuroSCORE*	French score	Magovern	New York State	Northern New England	Ontario	Parsonnet	Parsonnet (modified)	Pons	Toronto	Toronto (modified)	Tremblay	Tuman	UK national score	Veterans Affairs
Previous cardiac surgery	457 (7.3)	√		√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
Active endocarditis	55 (0.9)				√							√							
Heart failure	1156 (18.6)						√					√				√	√		√
Cardiomegaly	327 (5.3)						√												
Unstable angina	744 (12.0)		√		√							√				√			√
CCS class†	2.6 (1.0)																	√	
NYHA class†	2.4 (1.0)												√					√	√
Recent MI (within 24 hours)	144 (2.3)	√																	
Recent MI (within 48 hours)	207 (3.3)					√						√							
Recent MI (within 21 days)	793 (12.9)				√								√			√	√		
Ventricular arrhythmia (acute)	64 (1.0)					√		√				√							
Atrial fibrillation	508 (8.3)						√												
Pacemaker	33 (1.0)											√							
Left main stenosis	964 (17.9)											√		√	√			√	
Triple vessel disease	2690 (50.7)													√					
LV EF†	50 (12)	√		√	√	√	√	√	√	√	√	√		√	√	√	√	√	
Aortic gradient >120 mmHg	278 (4.5)										√	√							
Pulmonary hypertension	191 (3.1)				√						√	√					√		

The check mark indicates the risk variable included in each risk algorithm. *Additive and logistic. †Continuous variables are presented as mean (SD). The analysis is based on operations where the risk factor data was available.

Table 3.4c. Critical preoperative situations in 6222 open-heart operations.

Preoperative Risk Factors	n (%)	Amphiascore	Cabdeal	Cleveland Clinic	EuroSCORE*	French score	Magovern	New York State	Northern New England	Ontario	Parsonnet	Parsonnet (modified)	Pons	Toronto	Toronto (modified)	Tremblay	Tuman	UK national score	Veterans Affairs
Urgent surgery	1376 (22.2)						√		√	√				√	√			√	√
Emergency surgery	628 (10.1)	√	√	√	√		√		√	√			√	√	√	√	√	√	√
PTCA failure/complication	138 (2.2)						√				√	√							
Intubated	71 (1.1)					√						√	√					√	
IABP	134 (2.2)										√	√						√	√
Uncontrolled systemic disturbance†	1135 (18.2)															√			
Cardiogenic shock	78 (1.3)						√	√				√	√						
Hemodynamically unstable	286 (4.6)							√											
Critical state‡	308 (5.0)				√														
Catastrophic states§	206 (3.3)										√								

The check mark indicates the risk variable included in each risk algorithm. *Additive and logistic. †Any one or more of the following: systolic pulmonary arterial pressure >50 mmHg; uncontrolled systemic arterial hypertension; renal insufficiency; chronic lung disease; poor hepatic function; cerebrovascular insufficiency; severe arrhythmias; active endocarditis; cachexia. ‡Any one or more of the following: ventricular tachycardia or fibrillation or aborted sudden death; preoperative cardiac massage; preoperative ventilation before arrival in the operating room; preoperative inotropic support; intraaortic balloon counterpulsation; or preoperative acute renal failure (anuria or oliguria <10 ml/h). §Any one or more of the following: acute structural defect (acute ventricular septal defect or acute mitral valve regurgitation); cardiogenic shock; acute renal failure.

Table 3.4d. Surgical data in 6222 open-heart operations.

Surgical Procedures	n (%)	Amphiascore	Cabdeal	Cleveland Clinic	EuroSCORE*	French score	Magovern	New York State	Northern New England	Ontario	Parsonnet	Parsonnet (modified)	Pons	Toronto	Toronto (modified)	Tremblay	Tuman	UK national score	Veterans Affairs
Venous graft alone	572 (9.2)					√													
Single valve surgery only	657 (10.6)									√									
Valve surgery only	721 (11.6)																√		
Aortic valve surgery†	1106 (17.9)			√							√	√							
Mitral valve surgery‡	449 (7.3)	√		√							√	√	√						
Tricuspid valve surgery†	40 (0.6)					√						√	√						
Valve surgery and CABG	619 (9.9)					√				√	√	√	√				√		
Other than isolated CABG	1871 (30.1)				√														
Transplantation surgery	78 (1.3)					√													
Postinfarction septal rupture	37 (0.6)				√	√						√							
Left ventricular aneurysm	16 (0.3)										√	√	√						
Surgery on thoracic aorta	209 (3.4)				√								√						
Aortic dissection (acute)	79 (1.3)					√						√							

The check mark indicates the risk variable included in each risk algorithm. *Additive and logistic. †With or without CABG surgery. ‡With or without CABG surgery, except for Amphiascore where the definition is mitral valve surgery with CABG surgery.

Study III

Table 3.5. Patient characteristics in 3413 open-heart operations.

Variables	Mean±SD, or %
Age (years)	67.5±10.5
Female gender	27.8%
Chronic pulmonary disease	10.4%
Extracardiac arteriopathy	15.8%
Neurological dysfunction	5.9%
Previous cardiac surgery	4.5%
Serum creatinine >200μmol/l	2.4%
Active endocarditis	1.3%
Critical state (preoperative)#	7.0%
Unstable angina (requiring intravenous nitrates)	13.6%
Left ventricular dysfunction	
EF 30-50%	35.5%
EF <30%	9.0%
Preoperative myocardial infarction <90 days	34.4%
Pulmonary hypertension (systolic PAP >60 mmHg)	4.3%
Emergency	10.1%
Other than isolated CABG	27.1%
Surgery on thoracic aorta	3.5%
Postinfarction ventricular septal rupture closure	0.3%

For definition see Table 3.4c.

Study IV

Table 3.6a. Preoperative risk factors in 18362 open-heart operations.

Rank No	Risk Variables	Mean (SD) or n (%)
1	Age† (years)	62.6 (10.7)
2	One previous cardiac operation	1137 (6.2)
3	Left ventricular ejection fraction†	56 (15)
4	Serum creatinine† (μmol/l)	104 (49)
5	Emergency operation (<24 hours)	893 (4.9)
6	Acute aortic dissection	159 (0.9)
7	Thoracic aortic surgery	295 (1.6)
8	Heart or heart-lung transplantation	129 (0.7)
9	Aortic valve surgery for stenosis	2655 (14.5)
10	Acute active endocarditis	192 (1.0)
11	Urgent operation (stay in hospital before surgery)	3775 (20.6)
12	Mitral valve surgery for stenosis	946 (5.2)
13	Chronic congestive heart failure	1787 (9.7)
14	Intubated (before arrival in the operating room)	194 (1.1)
15	Carotid disease (unilateral stenosis >50%)	301 (1.6)
16	Intravenous inotropic support	425 (2.3)
17	Coronary bypass grafting	13286 (72.4)
18	Patient refusal of blood products	44 (0.2)
19	Atrial fibrillation	1676 (9.1)
20	Height† (cm)	168 (9)
21	Haematocrit† (%)	40 (4.8)
22	Long-term immunosuppressive therapy	76 (0.4)
23	Pulmonary embolectomy	14 (0.1)
24	Intra-aortic balloon pump	184 (1.0)
25	Previous surgery for vascular disease (carotids)	166 (0.9)
26	Intermittent claudication	1088 (5.9)
27	Systolic pulmonary artery pressure >60mm Hg	361 (2.0)
28	Tricuspid valve surgery	309 (1.7)
29	Postinfarction ventricular septal rupture closure	39 (0.2)
30	Neurological disorder	257 (1.4)
31	Cardiogenic shock	532 (2.9)
32	Mitral surgery for ischaemic acute regurgitation	49 (0.3)
33	No IMA (preoperative decision)	1480 (8.1)
34	Recent myocardial infarction† (number of days ago)	35 (25)

†Continuous variables are presented as mean (SD)

Table 3.6b. Preoperative risk factors in 18362 open-heart operations.

Rank No	Risk Variables	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† (mmHg)	76 (12)
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 two 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† (mmHg)	132 (20)
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 tumour known at surgery)	106 (0.6)
57	Mitral valve surgery for regurgitation	1671 (9.1)
58	Urine output <10ml/hour	137 (0.7)
59	Aortic valvular gradient >120 mmHg	215 (1.2)
60	Diabetes (diet-controlled)	1024 (5.6)
61	Chronic cardiac related dyspnoea at rest	1058 (5.8)
62	Chronic airway disease (treated)	726 (4.0)
63	Weight† (kg)	74 (13)
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)	80 (12)
72	Planned surgery for vascular disease (carotids)	85 (0.5)

†Continuous variables are presented as mean (SD)

Chapter 4

Statistical Methods

Means (SD) were used to describe continuous variables and frequencies were calculated for categorical variables.

Score-predicted operative mortality (death within 30 days of the operation) was calculated using the mean score for each risk model (*Study II*), except for the Northern New England algorithm where the published score-mortality table[24] was used.

One-way ANOVA was used to compare the difference between the predicted and the observed number of patients with an ICU stay >2 days for each risk cohort (*Study III*).

A proportion test[112] was used to compare the number of correctly classified patients by the ANN model versus the logistic EuroSCORE algorithm (*Study IV*).

4.1 Regression analyses and correlation tests

The Pearson correlation test was used to evaluate the correlation between the EuroSCORE and costs and LOS, respectively (*Study III*). Multivariable linear regression analysis was used to test which combinations of the individual risk factors in the EuroSCORE model that were significantly correlated to total cost. The cost and LOS were normalized by log-transforming the data[30].

Logistic regression analysis as described by Hosmer and Lemeshow[113] was performed to obtain the coefficients for the risk variables included in the logistic model used in *Study IV*.

4.2 Calibration

The EuroSCORE (*Study I and III*) and STS algorithms (*Study I*) were used as univariate predictors in developing logistic regression models to predict dichotomous outcomes such as mortality and ICU stay >2 days. The calibration of the algorithms was assessed by comparing the observed and the expected outcome for equal-sized quantiles of risk by using the Hosmer-Lemeshow goodness-of-fit test[46]. A p value >0.05 indicates a good accuracy.

4.3 Discrimination

The discriminatory power of the risk stratification models was evaluated by calculating the areas under ROC curves with 95% confidence limits[85]. An area of 1.0 under the ROC curve indicates perfect discrimination, whereas an area of 0.50 indicates complete absence of discrimination. Any intermediate value is a quantitative measure of the ability of the risk predictor model to distinguish between a positive and negative outcome, such as operative mortality and ICU stay shorter or longer than two days. To compare the areas under the resulting ROC curves the non-parametric approach described by DeLong and coworkers[114] was used. In *Study II*, the ROC area for each risk algorithm was systematically compared with the ROC area of the other 18 algorithms. The numbers of algorithms with a significantly larger or smaller ROC area was then computed. The probability significance level was adjusted for the effect of multiple comparisons using Sidak's method[87].

4.4 Training and validation of the ANN model

An ensemble approach was used, where several artificial neural networks were combined into a single prediction model. The individual members of the ensemble were standard MLP with one hidden layer and one output node that was used to encode the operative mortality[115]. Each MLP was trained using conjugate gradient descent applied to a mean square error function. To avoid over-

training and improve the generalization performance a weight decay regularization term was utilized (Table 1.2). 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 (i.e. the procedure to find the optimal set of model parameters and to select important risk variables) was performed using a six-fold cross-validation procedure (Figure 4.1).

The model selection procedure explored a large number of different parameter settings, by using high performance computer clusters, including the size of the MLP, weight decay parameters, size of the ensemble and risk factors identification (see below).

The final prediction models were tested on patients not previously exposed to the models, by using a 4-fold cross-testing technique. Thus, the patient material was randomly split into 4 groups. One of these groups was selected as the test set and excluded from further analysis. The remaining groups were used for training and validation. This procedure was performed 4 times with a new group selected each time as the test set (Figure 4.1).

4.5 Risk factor identification for mortality prediction

To select the most important risk variables and to minimize the number of variables included in the final ANN model, a ranking of risk variables was performed[52]. A baseline ROC area was created using all 72 variables. The ranking list was then obtained by measuring the change of the ROC area, as compared to 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.

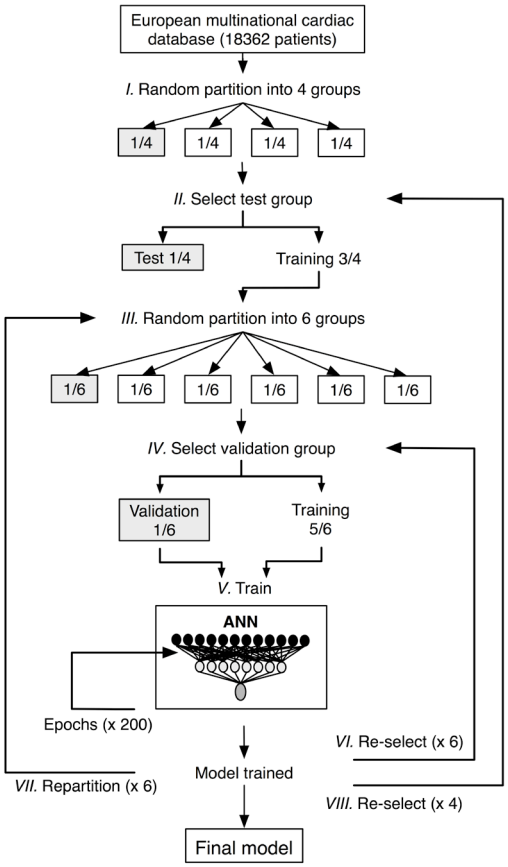


Figure 4.1. Schematic illustration of the ANN training and analysis process. The cardiac database was randomly split into 4 groups (I). One of these groups was selected as the test set and excluded from further analysis. The remaining groups were used for the training (II). Following the 6-fold cross 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 was optimized using 200 iterations (Figure 4.2). The procedure was repeated and a new validation group was selected (VI). After 6 re-selections all groups had been used for validation and the training group was repartitioned into 6 new groups (III) and the entire procedure were repeated. After 6 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.

4.6 Effective odds ratio and confidence intervals

The odds ratio for a specific risk variable in each patient (*Study IV)* was determined by changing the risk variable 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[70].

The 95% confidence intervals for both the output from the ANNs and the odds ratio were calculated using the bootstrap technique[54, 70].

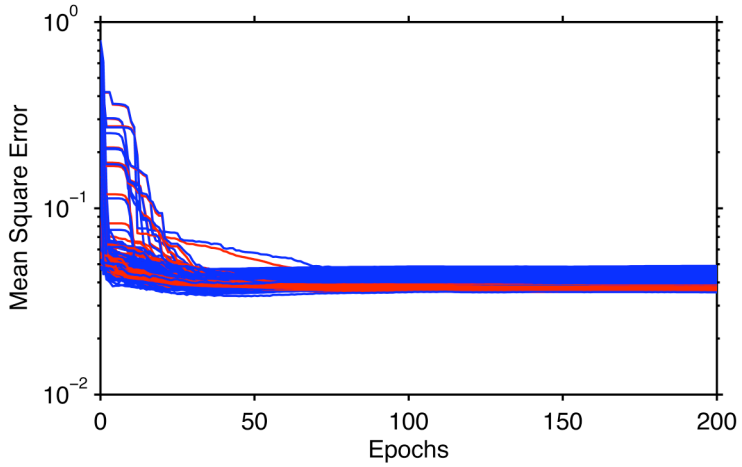


Figure 4.2. Monitoring the calibration of 144 ANNs. The average classification error per sample (using a summed square error function) is plotted (y-axis) during the training iterations (x-axis) for both the training and the validation samples. A pair of lines, red (training) and blue (validation), represents one model. The decrease in the classification errors with increasing epochs demonstrates the learning of the models to distinguish between survivors and non-survivors. There was no sign of over-training (over-fit): if the models begin to learn features in the training set, which are not present in the validation set, this would result in an increase in the error for the validation (blue lines) at that point, and the curves would diverge.

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 a classification (percentage mortality risk) for each individual patient, resulting in 1750 different classifications for each patient. Standard techniques[54, 70] were then used to extract the confidence intervals from these sets of risk predictions. The confidence intervals of the odds ratio for each risk variable were calculated in the same way.

4.7 Computer cluster and software

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 (Figure 4.3). The latter was also used for the statistical analysis.

The STS mortality risk was calculated by Summit Vista for Windows version 1.55 (1996), (Summit Medical Systems, Inc., Nice, France). The ANN calibration and analyses were performed with MatLab 7 (2005), Neural Network Toolbox (MathWorks, Natick, Massachusetts, USA). Graphs and statistical analyses were performed with Intercooled Stata version 9.0 (2005) statistical package, (StataCorp LP, College Station, Texas, USA).

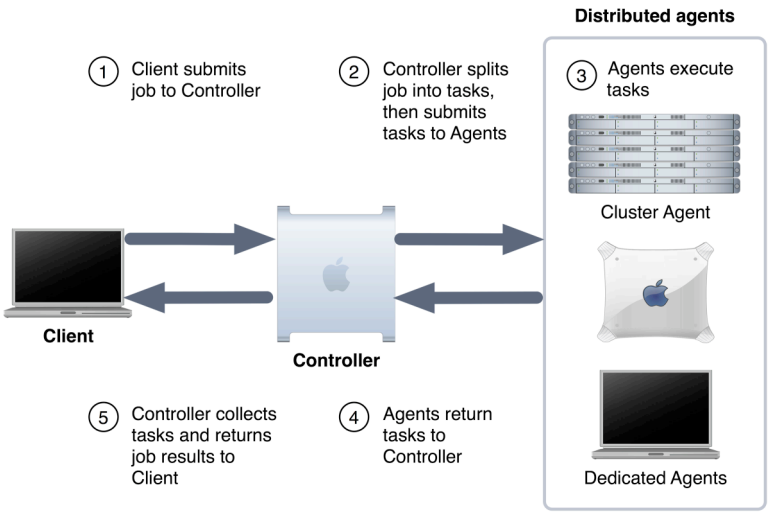


Figure 4.3. A schematic overview of the cluster software (Xgrid) in a Mac OS X cluster. One analysis (job) consists of 40 to 800 differently configured ANNs (1). Each configured ANN was submitted as a task to one of the seven agents by the controller (2). When all agents were occupied, the remaining task was put to a queue, awaiting the next available agent. The agent evaluated the ANN on the predefined patient group, which normally took 30-60 minutes (3). When finished, the task was returned to the controller (4). The controller collected all tasks and returned the final analysis (job) results to the client (5).

Chapter 5

Results

5.1 Study I

The study included 4497 CABG-only operations performed on 4487 patients. There was accurate documentation of vital status at 30 days in all cases. The average age was 66.4±9.3 years (range 31-90). The majority of patients were men (77.0%). The patient details are described in Table 3.3.

The actual 30-day mortality was 1.9%. The predicted versus the observed mortality for each risk score is plotted in Figure 5.1 (EuroSCORE) and Figure 5.2 (STS). The area under the ROC curve (Figure 5.3) was 0.84 (95% CI: 0.80 to 0.88) for EuroSCORE and 0.71 (95% CI: 0.66 to 0.77) for STS. The discriminatory power (area under the ROC curve) was significantly larger for EuroSCORE compared with STS ($p<0.0001$).

The study patients were grouped into ten different risk groups (as recommended by Hosmer and Lemeshow[46]) to test the calibration of the two models; however, because of ties the EuroSCORE has only nine distinct groups. The Hosmer-Lemeshow goodness-of-fit test with a *p* value of 0.81 for EuroSCORE and 0.83 for STS indicates a good accuracy of both models.

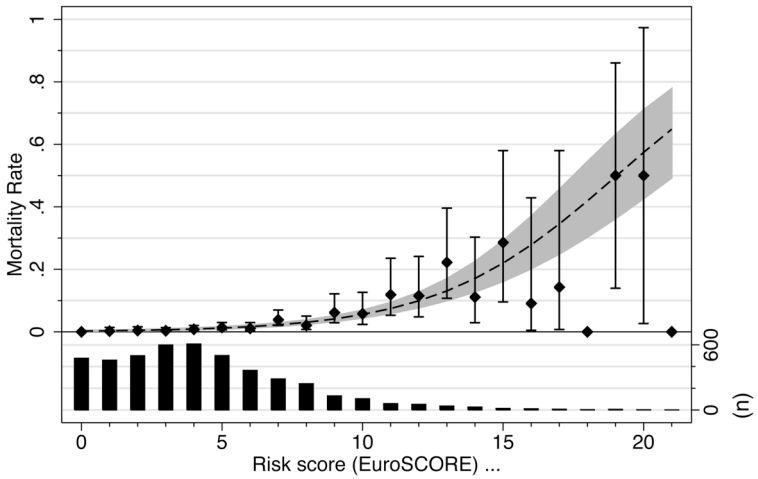


Figure 5.1. Graph of predicted versus observed 30-day mortality (left y-axis) for each EuroSCORE risk group. Predicted mortality (dotted line) with 95% CI (shadowed area); observed mortality (diamond) with 95% CI (bar). The histogram shows the number of patients (right y-axis) in each risk group.

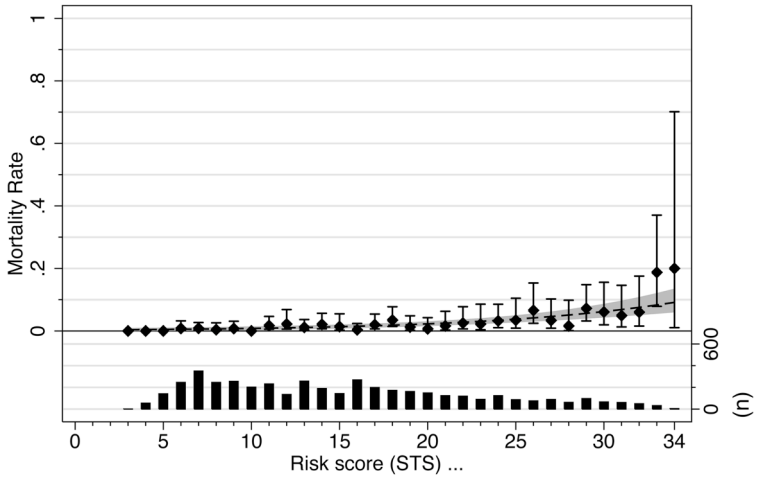


Figure 5.2. Graph of predicted versus observed 30-day mortality (left y-axis) for each STS risk group. Predicted mortality (dotted line) with 95% CI (shadowed area); observed mortality (diamond) with 95% CI (bar). The histogram shows the number of patients (right y-axis) in each risk group.

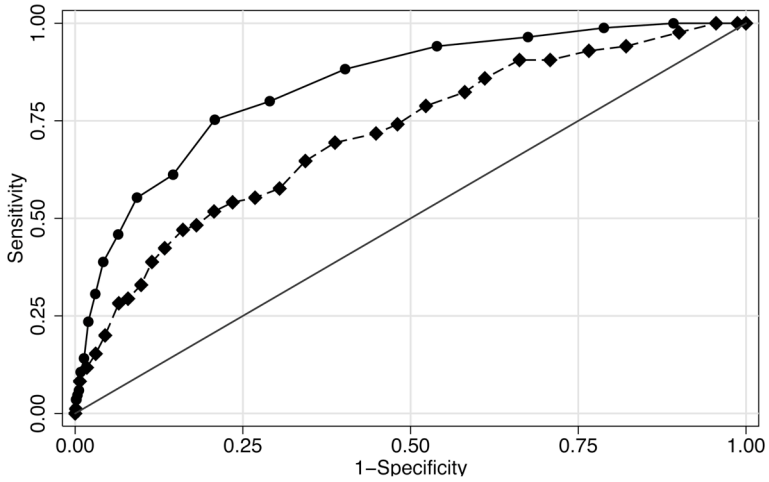


Figure 5.3. The ROC curve: The sensitivity of prediction of 30-day mortality is plotted versus the 1-specificity for the EuroSCORE (circles) and the STS (diamonds) risk stratification algorithm. The area under the curve for EuroSCORE is larger compared with STS. $\chi^2 = 22.90$; $p<0.0001$.

5.2 Study II

The study included 6222 open-heart operations performed on 6153 patients. In 2.0% of the total data points, missing values were replaced using the probability imputation technique[49]. There was accurate documentation of vital status and cause of death in all cases, and no patient was lost to follow-up. The average age was 66.3±10.6 years (range 18-95) and the majority of patients were men (71.6%). The patient details are described in Table 3.4a-d.

The discriminatory power (i.e. the area under the ROC curve) for 30-day mortality and one-year mortality was highest for the logistic (0.84 and 0.77) and additive (0.84 and 0.77) EuroSCORE algorithms, followed by the Cleveland Clinic (0.82 and 0.76) and the Magovern (0.82 and 0.76) scoring systems. None of the other risk algorithms had a significantly better discriminatory power (larger ROC area) than these four (Figure 5.4). In the sub-analysis with CABG-only patients the discriminatory power of the two EuroSCORE algorithms were highest, followed by the NYS and Cleveland Clinic risk algorithms (Table 5.1).

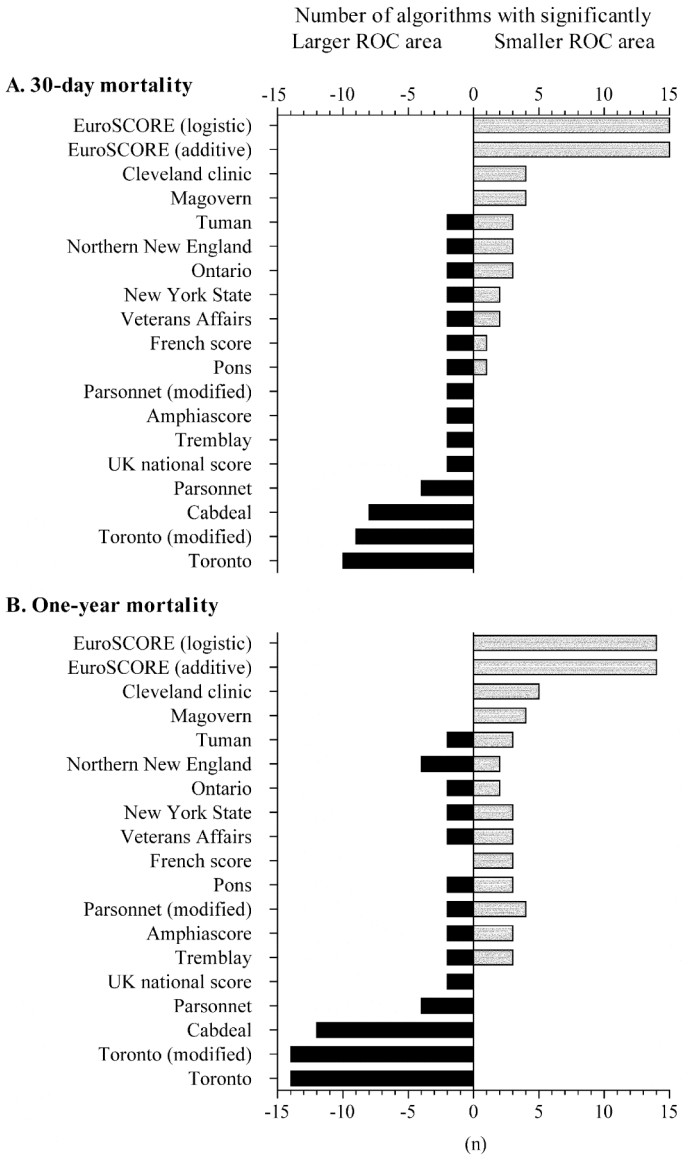


Figure 5.4. Comparison of the ROC area for different risk algorithms. For each risk scoring system (left y-axis), the number of risk algorithms with a significantly (p<0.05) larger (black bar) or smaller (grey bar) ROC area are shown. A: 30-day mortality and B: one-year mortality. Open-heart surgery (n=6222).

Table 5.1. ROC area for the five risk algorithms with best performance and accuracy in CABG-only surgery (n=4351).

	30-day mortality ROC area (95% CI)		One-year mortality ROC area (95% CI)	
EuroSCORE (logistic)	0.86	(0.82-0.90)	0.75	(0.72-0.79)
EuroSCORE (additive)	0.85	(0.81-0.89)	0.75	(0.71-0.78)
New York State	0.84	(0.80-0.88)	0.75	(0.72-0.79)
Cleveland Clinic	0.84	(0.80-0.88)	0.75	(0.71-0.78)
Parsonnet (modified)	0.84	(0.80-0.88)	0.73	(0.69-0.77)

The most common cause of death within 30 days was cardiovascular disease (n=163, 91%), followed by cerebrovascular disease (n=3, 1.7%) and malignant neoplasm (n=3, 1.7%). Cardiovascular disease was also the most common cause of death within one year (n=280, 74%) (Table 5.2). For each risk algorithm, the ROC areas for cardiovascular-related (n=163) and total 30-day mortality (n=180) were almost identical (difference 0.005 or less). The discriminatory power for cardiovascular-related one-year mortality (n=280) increased by approximately 0.03 for all 19 algorithms compared with the discriminatory power for total one-year mortality (n=377). However, it did not change their relative order of discriminatory power.

Table 5.2. Causes of death.

	30-day mortality		One-year mortality	
	n	(%)	n	(%)
Cardiovascular disease	163	(90.6)	280	(74.3)
Cerebrovascular disease	3	(1.7)	16	(4.2)
Chronic lower respiratory disease	2	(1.1)	10	(2.7)
Diabetes mellitus	0	0	6	(1.6)
Malignant neoplasm	3	(1.7)	22	(5.8)
Miscellaneous	8	(4.4)	31	(8.2)
Renal disease	0	0	2	(0.5)
Septicaemia	1	(0.6)	10	(2.7)

5.3 Study III

The study included 3413 open-heart operations performed on 3404 patients. There was accurate documentation of data including 30-day mortality in all cases. The average age was 67.5±10.5 years (range 18 to 89 years). The majority of patients were men (72.2%). The patient details are described in Table 3.5. A CABG-only operation was performed in 2487 cases (72.9%), 710 (20.8%) cases had a valve procedure with or without CABG surgery, and 216 (6.3%) were miscellaneous procedures (post-infarction septal rupture, aortic aneurysm or dissection, etc).

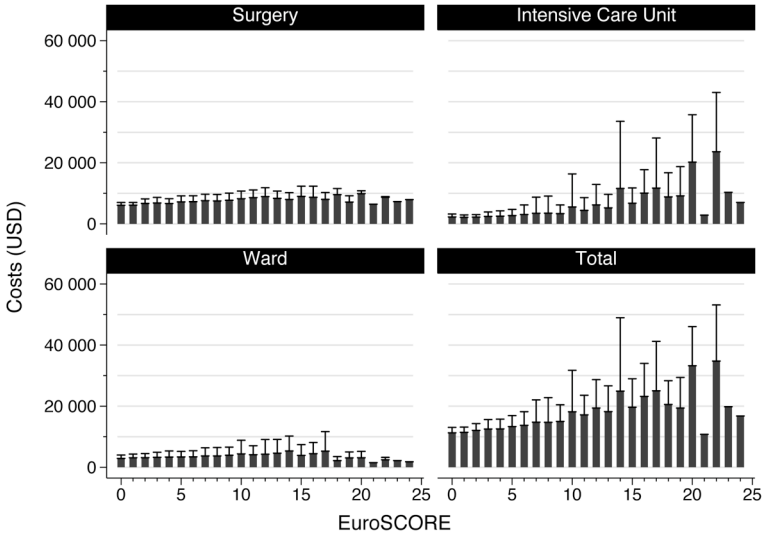


Figure 5.5. Graph of costs (mean ±SD, bar) for each risk score.

The actual 30-day postoperative mortality was 2.5%. The mean cost for the surgery was 7300 ± 2120 USD (median 6613 USD, range 2563 to 25988 USD), in the ICU 3746 ± 6032 USD (median 2182 USD, range 632 to 134263 USD), in the ward 3500 ± 2605 USD (median 2999 USD, range 0 to 41626 USD) and the mean total cost was 14546 ±7658 USD (median 12546 USD, range 6995 to 157912 USD). The mean costs (±SD) were calculated for the EuroSCORE risk groups 0 to 24 for the surgery, the ICU and the ward, with the results shown in Figure 5.5. The log-transformed cost for the individual patients was significantly correlated to EuroSCORE. The strongest correlation was

between the EuroSCORE and log-transformed ICU costs, with a correlation coefficient (r) of 0.46 (p<0.0001) (Table 5.3).

Table 5.3. Regression analysis results (n=3413).

Comparison	Equation	Coefficient 95% CI	r	p	r^2
Cost versus EuroSCORE					
Surgery	6182 USD x $1.021^{\text{EuroSCORE}}$	1.019-1.024	0.31	<0.0001	0.10
ICU	1752 USD x $1.076^{\text{EuroSCORE}}$	1.071-1.081	0.46	<0.0001	0.21
Ward	2873 USD x $1.014^{\text{EuroSCORE}}$	1.010-1.019	0.11	<0.0001	0.01
Total	10688 USD x $1.040^{\text{EuroSCORE}}$	1.038-1.043	0.47	<0.0001	0.22
LOS versus EuroSCORE					
ICU	0.88 days x $1.071^{\text{EuroSCORE}}$	1.066-1.076	0.45	<0.0001	0.21
Ward	6.05 days x $1.015^{\text{EuroSCORE}}$	1.010-1.020	0.11	<0.0001	0.01
Total	7.14 days x $1.028^{\text{EuroSCORE}}$	1.024-1.032	0.24	<0.0001	0.06
Probability of ICU stay >1 day	$\frac{e^{(-3.0+0.28 \times \text{EuroSCORE})}}{1 + e^{(-3.0+0.28 \times \text{EuroSCORE})}}$	0.25-0.30		<0.0001	0.16#
Probability of ICU stay >2 days	$\frac{e^{(-4.0+0.29 \times \text{EuroSCORE})}}{1 + e^{(-4.0+0.29 \times \text{EuroSCORE})}}$	0.26-0.32		<0.0001	0.18#

pseudo r^2

When patients were grouped in cohorts of similar predicted EuroSCORE risk (Table 3.2) the correlation between log-transformed mean costs was improved. The mean total cost was significantly correlated to mean EuroSCORE risk for each risk cohort, with a correlation coefficient (r) of 0.99 (p<0.0001); r was 0.99 for the mean surgery cost, 0.98 for the mean ICU cost, and 0.94 for the mean ward cost.

In the multivariable linear regression analysis with the 18 EuroSCORE risk factors as regressor variables and log-transformed cost as the dependent variable, 15 EuroSCORE variables were found to be significantly (p<0.05) associated with the log-transformed cost (Table 5.4) with a correlation coefficient (r) of 0.63 (p<0.0001).

The mean LOS in the ICU was 1.76±2.39 days, (median 1 day, range 1 to 41 days). Log-transformed LOS at the ICU was significantly correlated to EuroSCORE with a correlation coefficient

(r) of 0.45 (p<0.0001) (Table 5.3). Of all patients, 13.7% had an ICU stay >2 days. Hosmer-Lemeshow test yielded a p value of 0.40 for the EuroSCORE predicting an ICU stay >2 days, which indicates a good accuracy of the model. The area under the ROC curve for an ICU stay >2 days was 0.78 (95% CI: 0.76 to 0.81). The probability of an ICU stay exceeding 2 days was >50% at a EuroSCORE of 14 or more (Figure 5.6). The sensitivity and specificity at this cut-off point was 21% and 98%, respectively.

During the entire study period (169 weeks), the mean weekly number of patients entering the ICU was 20±7.13 (median 22, range 5-35). During this period, the EuroSCORE algorithm predicted the number of patients with an ICU stay >2 days exactly in 51 weeks (30%), and within ±1 patient in 127 weeks (75%). The predictive accuracy was independent of the EuroSCORE risk cohort (p=0.65).

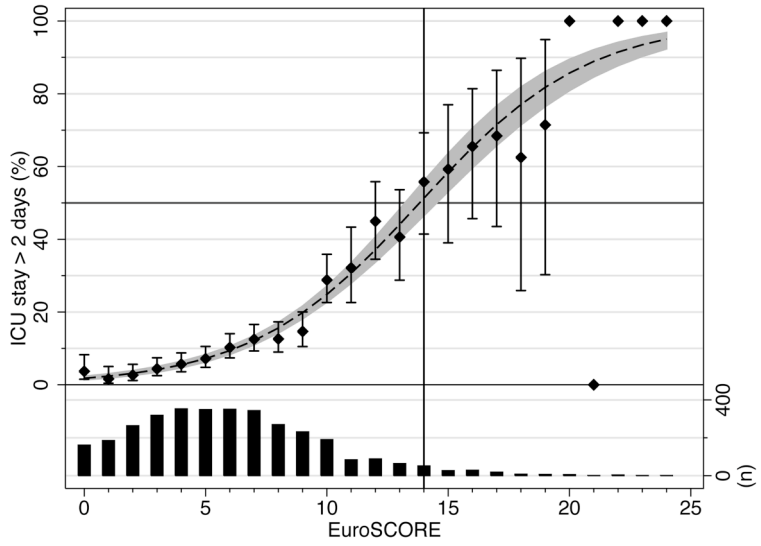


Figure 5.6. Percentage of patients with an ICU stay >2 days (left y-axis) for each EuroSCORE risk group (x-axis). Predicted ICU stay >2 days (dotted line) with 95% confidence interval (shadowed area); observed ICU stay >2 days (diamond) with 95% confidence interval (bar). The histogram shows the number of patients (right y-axis) in each risk group.

Table 5.4. Regression analysis: EuroSCORE variables independently associated with total cost (n=3413).

Variables	Cost increase (%)	95% CI	*p*
Age*	0.28	0.20- 0.37	<0.0001
Female gender	-1.1	-3.0- 0.89	0.283
Chronic pulmonary disease	2.9	0.0- 5.8	0.047
Extracardiac arteriopathy	3.4	1.0- 6.0	0.006
Neurological dysfunction	6.9	3.1- 10.9	<0.0001
Previous cardiac surgery	12.0	7.5- 16.8	<0.0001
Serum creatinine >200μmol/l (2.27mg/dL)	16.4	10.0- 23.2	<0.0001
Active endocarditis	11.4	3.1- 20.4	0.007
Critical preoperative state#	19.9	14.9- 25.1	<0.0001
Unstable angina	2.2	-0.7- 5.3	0.132
EF 30-50%	3.3	1.4- 5.3	0.001
EF <30%	12.1	8.6- 15.8	<0.0001
Preoperative MI <90 days	2.7	0.6- 4.7	0.010
Pulmonary hypertension	15.5	10.5- 20.8	<0.0001
Emergency	9.3	5.3- 13.5	<0.0001
Other than isolated CABG	37.5	34.4- 40.6	<0.0001
Surgery on thoracic aorta	24.4	18.2- 30.8	<0.0001
Postinfarction ventricular septal rupture closure	12.0	-4.9- 31.9	0.174

* % increase per year of age. # For definition see Table 3.4c

5.4 Study IV

From the EuroSCORE database 18362 patients were included in the analysis. In 0.7% of the total data points missing values were imputed, as described. The average age was 62.6±10.7 years (range 17-89) and the majority of patients were men (71.7%). The patient details are described in Table 3.6a-b. The actual operative mortality was 4.9% (n=891).

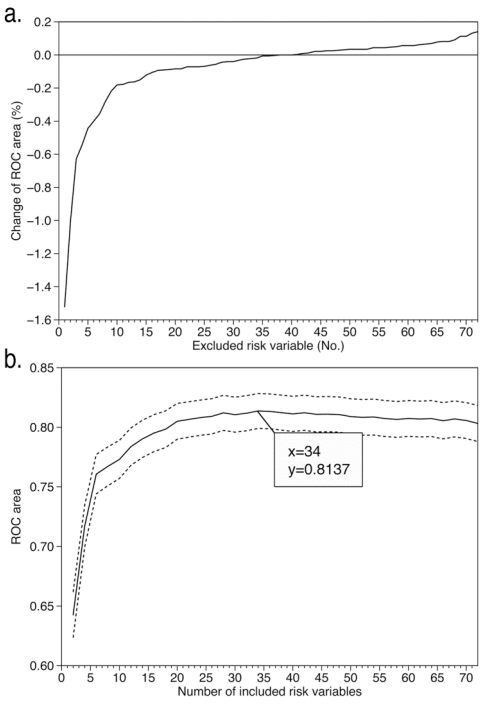


Figure 5.7.

a. The graph shows the difference (%) in the validation ROC area (y-axis) from each ANN model including 71 risk variables, compared with the model including all 72 risk variables. The x-axis shows the excluded risk variable number (No.), in order of importance (see Tables 3.6a-b).

b. The solid line shows the validation ROC area (y-axis) from the ANNs with different numbers of included risk variables (x-axis). Dashed lines indicate 95 % CI.

Approximately 42500 different ANN models were validated. The architecture for the final validation ANN model included one hidden layer with 14 nodes, one 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. The largest validation ROC area, 0.82 (95% CI: 0.80 to 0.83) was achieved when 34 of the top ranked risk variables were included (Figure 5.7). 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 eight nodes in the

hidden layer. All artificial networks from the six-fold cross-validation procedure were saved, resulting in 36 individual networks from the six members in the ensemble. Thus, an ensemble size of 36 was used to classify the patients in the test sets. The discriminatory power (i.e. the area under the ROC curve) for operative mortality was significantly larger for the final ANNs, 0.81 (95% CI: 0.79 to 0.82) compared with the logistic EuroSCORE model, 0.79 (95% CI: 0.78 to 0.81), p=0.0001 (Figure 5.8). 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 to 0.81), p<0.0001. The numbers of correctly classified survivors for a sensitivity of 25%, 50% and 75%, were 17051, 15577 and 12438 patients for the ANNs, and 16990, 15321 and 11718 patients for the logistic EuroSCORE. The difference between the ANNs and logistic EuroSCORE was significant for all three sensitivity cut-off values: p=0.0395, p<0.0001 and p<0.0001, respectively. For the different surgical procedures (CABG-only, valve procedure with or without CABG surgery and miscellaneous procedures) there were no differences in discriminatory power for the ANNs, but there were significant differences for the logistic EuroSCORE (Table 5.5).

Table 5.5. The ROC area from the test data set for different surgical procedures.

Surgical procedure	n	ROC area ANNs	ROC area Logistic EuroSCORE
CABG-only surgery	11 628	0.80 (0.77-0.82)	0.78 (0.75-0.80)
Valve procedure*	4 907	0.76 (0.73-0.79)	0.72 (0.69-0.75)
Miscellaneous	1 827	0.80 (0.77-0.83)	0.78 (0.75-0.82)
p value		0.15	0.0072

* Valve procedure with or without CABG surgery.

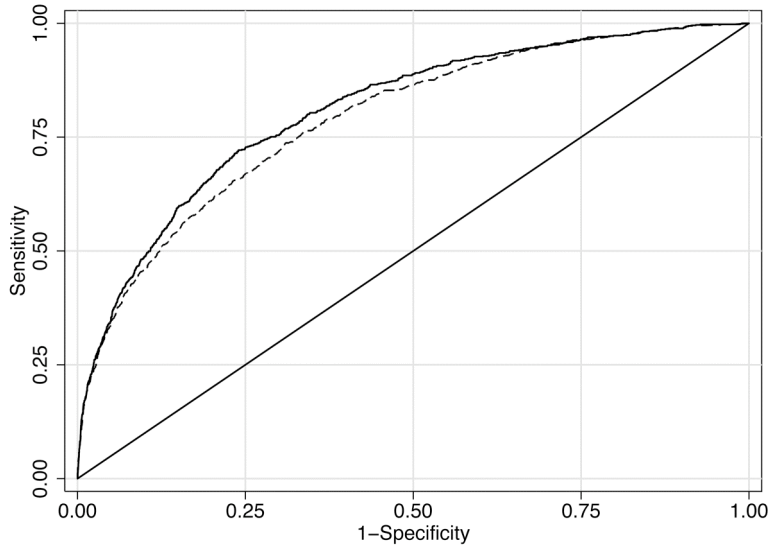


Figure 5.8. The receiver operating characteristic curves (ROC) from the test data set: The ANNs (solid line) and the logistic EuroSCORE (dashed line) risk stratification algorithms. The area under the curve for ANNs is larger compared with the logistic EuroSCORE. $\chi^2 = 15.7$; p=0.0001.

Bootstrap sampling was used to generate confidence intervals for the ANN classification (Figure 5.9). An individual patient with a calculated mortality risk of 0.64 belongs with at least 95% certainty to the group of patients not likely to survive the operation (i.e. a mortality risk exceeding 50%). For a patient with a calculated risk of 0.31 the opposite holds true.

To evaluate if the final ANN risk prediction model was applicable to a patient cohort which had not been used in the development of the ANNs or participated in the EuroSCORE project, a subset (n=1246) from a local database (*Database I*) 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.

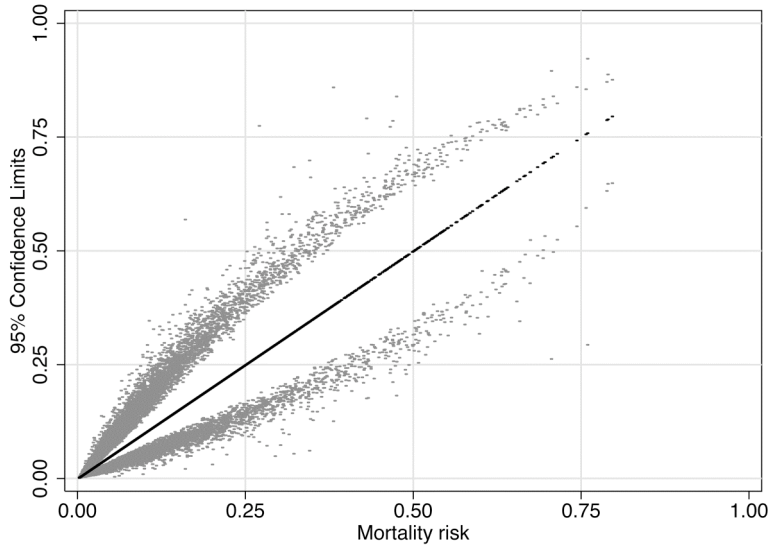


Figure 5.9. Graph of predicted mortality risk (x-axis) (black dots) and the 95% confidence limits (y-axis) (grey dots) calculated by the ANNs for each individual patient.

Chapter 6

General Discussion

Preoperative evaluation of a patient's surgical risk is an important component in cardiac surgery. Risk stratification can provide insight into the risk of complications and mortality, guide the selection of patients for surgery versus alternative therapies[12, 13, 24], and may also predict the need for hospital care resources[28, 29, 31] and improve the quality of care[7, 8].

6.1 Performance and accuracy of risk score systems

The search for an effective method of mortality prediction in open-heart surgery started in the 1980s[19]. Several risk score algorithms for cardiac surgery have been published (Table 1.3). A few comparative studies of different risk algorithms exist, but the relative performance of the risk scoring systems currently used has been unclear.

The potential of ROC curves in medical diagnostic testing was recognized as early as in 1960[83]. Even if comparison of ROC curves to evaluate models in a statistically valid fashion remains controversial, the ROC curve is currently the best developed statistical tool for describing performance[116].

The first study in this thesis aimed to compare the accuracy and performance of the STS and the EuroSCORE risk stratification algorithms in a CABG-only population. The EuroSCORE model has previously been shown to work well across many European countries[117] and in North America[118]. EuroSCORE has also been shown to perform well in comparison with other risk algorithms such

as Cleveland Clinic[26], Parsonnet[19] and French Score[90]. However, comparative studies of the EuroSCORE[60] and STS[11] scoring systems has hitherto been lacking. The more complicated STS database algorithms remain proprietary and confidential, which could explain why only a few studies comparing the STS database algorithm with other risk algorithms have been published[10, 119].

In the present study, a superior discriminatory power was achieved using the EuroSCORE algorithm. The obtained ROC-area of 0.71 for the STS is comparable to the results from earlier studies (0.64-0.81)[10, 119], which corroborate the results from *Study I*. On the other hand, the predictive accuracy of the two risk scoring systems may be influenced by the different time periods in which they were developed.

In earlier studies comparing mortality prediction by different algorithms, no significant differences in performance have been found[89-91]. This may be explained by small patient study groups[90, 91]. In *Study II*, four risk scoring systems (the two EuroSCORE algorithms, Cleveland Clinic and Magovern risk score) had a significantly better performance than the other 15 algorithms. Compared with the earlier mentioned studies, a relatively large patient population was used in *Study II,* which may have made it easier to identify the superior algorithms. The systematic comparison methodology used, which was not utilized in the other studies, may be another explanation. An additional finding in *Study II* was also that the algorithms could be used to predict long-term (one-year) mortality, especially for cardiovascular deaths.

6.2 Prediction of resource utilization

Earlier studies have indicated that preoperative risk variables can be used to predict costs of cardiac surgery[120, 121]. Fifteen of the 18 EuroSCORE variables were found to be significantly correlated to cost of care in *Study III*. An increasing Cleveland Clinic risk score has previously been shown to be associated with an increase in total cost and longer postoperative LOS[122]. Similar results has been shown with the Cabdeal risk algorithm[123]. Riordan et al[29] found that grouping patients in risk cohorts resulted in a correlation between the STS risk

algorithm and total cost, but for individual patients the prediction was poor.

The results from *Study I and II* make the use of the EuroSCORE risk algorithm to predict the need for different resources logical. Pintor et al[30] and Sokolovic et al[31] recently demonstrated a correlation between EuroSCORE and cost of open-heart surgery, similar to the results found in *Study III*. However, these studies were rather small (488 and 201 patients, respectively) and focused on total cost and not on the different components of heart surgery resource utilization (surgery, ICU and ward), as in *Study III*.

Several studies have attempted to identify preoperative risk variables that predict LOS in the ICU following cardiac surgery. The Parsonnet risk algorithm seems able to preoperatively identify patients likely to spend more than 24 hours in the ICU[124]. In that study, the discriminatory power (ROC area) of 0.70 was less than in *Study III* (0.76 for ICU stay >1 day and 0.78 for ICU stay >2 days). Three additional studies[93, 125, 126], comparing different risk algorithms, have also found a correlation between EuroSCORE and ICU stay >2 days. In our experience, patients staying at the ICU >2 days are likely to remain there for prolonged periods. Like other groups[93, 125, 126], we therefore chose to focus the additional analyses on >2 day ICU stays, as being clinically more relevant.

Strengths of *Study I-III* are that the algorithms could be analysed using a relatively large patient material where the patient data was collected on a regular basis in the daily clinical work, and that the data was preoperatively entered into the database, generally by residents and not by the surgeon performing the operation.

6.3 Artificial neural networks

Different methods to improve the accuracy of risk algorithms have been suggested, e.g. to include more patients with higher risk, to 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[27, 45, 127]. Tu and Guerriere[128] used a neural network as a predictive instrument for ICU stay, finding both

advantages and disadvantages compared with other statistical techniques. Only a few studies have investigated ANNs in the prediction of survival after cardiac surgery[69, 70, 75-78]. Most of these are based on CABG-only patients[69, 70, 76-78] and only one included all cardiac surgical procedures[75]. None of these studies found any considerable improvement over the traditional bio-statistical methods. Orr *et al* [75] and Tu *et al* [69] showed that an ANN could be used to estimate cardiac surgical mortality, but the performance was equivalent to that of logistic regression. These two studies were made on smaller cohorts than *Study IV* (1477 and 4782 patients) and used a limited number of risk variables (7 and 11). Lippmann *et al* [70] obtained a similar result when ANNs were used on patients from the STS database. Despite that 80000 patients with 32 risk variables were included in the study, the ANNs showed a performance equivalent to the other prediction models.

The differences between the studies mentioned and *Study IV* are several. The ANN model in *Study IV* was developed on a large multi-institutional database from eight European countries, and the patient data was quality-checked and validated by two independent operators before it was entered into the database[110]. No prior variable selection was used; instead all available variables were initially included, and the most important were obtained by a variable ranking and minimizing procedure. No categorization was performed, and the ANN architecture was achieved by exploring a large number of configurations (42500) by using high performance computer clusters. An aim in *Study IV* was to avoid most of the limitations discovered in the earlier ANN studies, such as variable categorization and risk factor identification by using traditional significance testing[52].

6.4 Risk factor identification for mortality prediction

One fundamental and controversial question is the number of variables optimally included in a risk model. As mentioned before, too many variables may lead to over-fitting of the model, instability, increased cost and difficulties of data collection. Too few variables, on the other hand, may decrease the performance of the model. In *Study II*, the analysed 19 risk score models included 6 to 33 risk variables.

To evaluate if the performance was correlated to the number of risk variables in each model, a non-linear regression analysis was performed (Figure 6.1). Even if there is no statistical certainty in this analysis, it suggests that too few and too many variables may decrease the model performance.

In almost all of the investigated algorithms (both in earlier published studies and those presented here) all variables are categorical. Continuous data are preferable, to avoid arbitrariness and loss of valuable information that may occur with categorization. Identifying a nonlinear relationship is also more likely using continuous rather than categorical variables[129]. In S*tudy IV* a total of 72 variables (11 continuous) were used. No prior variable selection such as significance testing was performed; 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, and was significantly larger compared with a model including all 72 variables. However, the discriminatory power (ROC area) for a logistic regression model including the same 34 variables was significantly lower than the ROC area for the final ANN model. The results from *Study IV* thus indicate that both the ANN technique and the variable selection process are important for optimal outcome prediction.

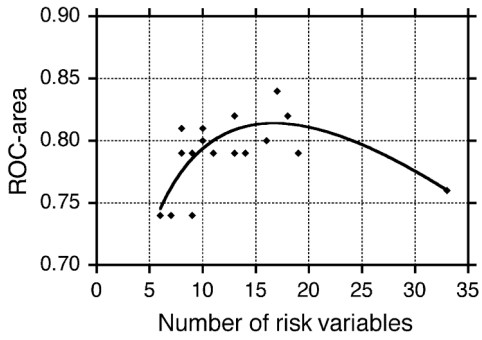


Figure 6.1. The ROC area (y-axis) corresponding to the number of included risk variables (x-axis) for different risk score algorithms (Table 1.3). Goodness-of-fit $r^2 = 0.52$.

6.5 Inaccuracy of individual outcome prediction

Most algorithms overestimated the 30-day mortality in the investigated patient population (*Study I and II*). The same finding has been reported in other studies[89, 90]. Rather than reflecting weaknesses in the risk score algorithm, these findings are probably explained by differences in patient mix and temporal periods, compared with the original databases used for development of the algorithms[90]. Prediction of mortality rate in the CABG-only patients was almost perfect using the NNE and NYS algorithms, which are both newly developed and for use in CABG-only surgery.

The EuroSCORE predictive performance of resource utilization was less at higher risk scores than at lower scores. This pattern has previously been reported in mortality studies[32] and was also seen in *Study I*. The lower numbers of patients in high risk score groups might contribute to these findings. *Study III* also showed, as in other investigations[29, 130], that the predictive value was limited for individual patients, but an excellent correlation was seen if the patients were grouped in risk cohorts. To improve the prediction for the individual patient, a machine-learning technique with the bootstrap method was used in *Study IV* to generate individual confidence intervals (Figure 5.9). These increase the accuracy of the mortality prediction. This kind of analysis is not common, but Michel et al[131] recently published a similar analysis for the logistic EuroSCORE model. Even if the risk stratification for an individual patient has thus become more accurate compared with earlier developed algorithms, there is still room for improvement.

6.6 Factors influencing accuracy

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, geographical differences in patient risk factors, inclusion and surgical procedure selection criteria, and gaming. The prevalence of risk factors in patients referred for heart surgery may also change over time.

6.6.1 Variable frequency and definition

When comparison of the accuracy and predictive power of large databases are attempted, such as in *Study II,* difficulties arise regarding variable definitions and the prevalence of risk factors. One weakness of multivariable regression techniques is that some variables may have too low frequencies to be used in a multivariable regression model, even if they contribute significantly to the outcome[18]. An advantage of ANNs are that they use computer iteration to identify a pattern of variables associated with outcome, and are far less affected by low frequencies in particular variables[69, 70]. However, ROC analysis is also a robust technique for such comparisons. Murphy-Filkins et al[132] showed that an increase of up to five times of a low-frequency variable (for example due to a difference in a variable definition) did not appreciably change the model fit.

The patient form used in *Study I* was constructed in 1996, three years before the EuroSCORE was officially published. For this reason, the definition of one of the risk variables was not identical with the definition from the original EuroSCORE publication, as preoperative myocardial infarction was defined to be present within 21 days before the operation (STS definition) rather than within 90 days preoperatively (EuroSCORE definition). To evaluate if the change of definition may have influenced our overall results, we performed separate calculations for the patients included in the database after this date (n=1130), applying both definitions for preoperative myocardial infarction. The area under the ROC curve remained significantly larger for the EuroSCORE algorithms compared with STS, regardless of the definition of preoperative myocardial infarction used.

6.6.2 Incomplete data fields

All databases may have incomplete data fields, especially if a large number of risk factors are included such as in *Studies II and IV*. The ANN model is tolerant of missing data[70], but risk score systems built upon logistic regression analysis are not[43]. This makes data imputation techniques necessary. The probability imputation technique, used in *Study II*, has been shown to work well in prognostic factor studies[51]. There are, however, several other techniques to handle missing values[43]. The most common method is to exclude the patients with

missing values from analysis. Because missing values are more likely in emergent high-risk patients, this could result in bias.

In *Study II* several techniques were explored. The results of these were similar, except when the patients with incomplete data were excluded, which altered the relative performance of some of the low ranked algorithms. However, for the top ranked algorithms there were no differences in the relative order. The probability imputation technique was chosen because of its simplicity and documentation[49, 51].

6.6.3 Geographical differences in patient risk factors

The geographical differences in the occurrence of patient risk factors may have influenced the design of different risk scoring systems. However, it does not seem to influence the present results. The best-performing risk score systems in *Study II* were developed in two different geographical areas: Europe and the US. Nashef et al have also demonstrated that the EuroSCORE works well when applied to a Northern American patient population[118].

6.6.4 Surgical procedure

Eight of the risk algorithms included in *Study II* (Cabdeal, NYS, NNE, Magovern, Toronto, Toronto (modified), UK National Score and VA) were originally designed to predict early mortality in CABG-only patients, which could affect the predictive accuracy. A sub-analysis of CABG-only patients was therefore performed. The same two risk-scoring systems, logistic and additive EuroSCORE, were identified as having superior performance, followed by the NYS and the Cleveland Clinic risk-scoring systems.

Earlier studies on risk analysis in cardiac surgery have mostly been developed and validated on CABG-only patients[7, 24] or all open-heart surgery procedures[110]. Recently, the NNE Cardiovascular Disease Study Group presented a risk model for aortic valve surgery and one for mitral valve surgery[104]. Analyses comparing risk score performance in different surgical procedures have been lacking. In *Study IV* this analysis was performed. The ANN model showed similar performance independent of the surgical procedure, unlike the logistic EuroSCORE model. This may be explained by a better risk factor selection by the ANN, but also by the capacity of the ANN model to recognize complex nonlinear relationships.

6.6.5 Inclusion criteria

In *Study III* cases dying intra-operatively were excluded, since they did not require any postoperative resources. This could be debated: The exclusion of these mainly high-risk patients probably reduces the actual predictive power of the analysis, but the differences will be minor since the number of patients who died intra-operatively was small in this study (1.1% during the period in question).

6.6.6 Change in risk factor prevalence over time

The prevalence of risk factors in patients referred for heart surgery may show temporal changes, which could be a limitation in *Studies I, II and IV*.

The STS risk analysis algorithm in *Study I* is based on the STS results from the years 1990-1993 (version 2.0), whereas the EuroSCORE algorithm was developed approximately five years later. To evaluate if the risk algorithm performance changed over time in the local database, the risk score ROC areas for each year from 1996 to 2000 were compared. There was no significant difference in performance or any apparent trend over time (Table 6.1a-b). Of course, the sample sizes became smaller, making it more difficult to identify a difference.

Table 6.1a. The ROC area of the EuroSCORE algorithm over time (CABG-only surgery).

Year	Procedures (n)	ROC area	95% confidence interval
1996	957	0.88	0.81-0.95
1997	1049	0.87	0.76-0.98
1998	752	0.79	0.66-0.92
1999	833	0.78	0.65-0.91
2000	752	0.82	0.73-0.90

Chi square test: $\chi^2 = 3.26$, $p = 0.516$

Table 6.1b. The ROC area of the STS algorithm over time (CABG-only surgery).

Year	Procedures (n)	ROC area	95% confidence interval
1996	957	0.69	0.58-0.80
1997	1049	0.77	0.61-0.93
1998	752	0.70	0.49-0.91
1999	833	0.64	0.46-0.82
2000	752	0.72	0.62-0.82

Chi square test: $\chi^2 = 1.41$, $p = 0.843$

The same limitation is possible in *Study IV*, which was performed on data collected ten years ago. However, a similar result was obtained in the blind test (using the local database), where the surgical procedures were performed between 1996 and 2001.

6.6.7 Gaming

In the clinical use of risk algorithms, patients with higher risk will have higher predicted mortality. In other words, the sicker the patient population is, the higher the acceptable operative mortality will be. This could incite exaggeration of preoperative risk factor severity. Several definitions of risk factors are open to some degree of clinical interpretation. If the preoperative recording of risk factors is thus subject to bias, the information in the database may suffer from "clinical inflation", and the risk stratification model will be less accurate[63].

6.7 Future perspectives

The future will certainly bring refinements of risk adjustment methods and increasing use of risk evaluation in many areas of health care delivery. There will probably be an evolution from models focused primarily on operative mortality to an increased awareness of other endpoints. Postoperative complications, readmissions, functional status and quality of life measurements will be more common endpoints in the future, and long-term results will be emphasized. Future risk stratification models may employ novel biochemical, physiological or genetic risk markers and more complex and computer-intensive algorithms such as machine learning, as high-performance computer clusters become more available.

These more complicated risk models can not compete with the additive algorithms when it comes to simplicity, but can be made available on the hospital intranet or publicly on the Internet, as exemplified by the logistic EuroSCORE risk model (http://www.euroscore.org/calc.html) and the STS Risk Calculator (http://www.sts.org/sections/stsnationaldatabase/riskcalculator/).

With the use of modern information technology systems and computerized medical records, the risk factors for a patient referred for a specific treatment may be available on-line. Depending on the present risk factors, a waiting list score, mortality scores for different treatments (e.g. cardiac surgery or PTCA) and a postoperative morbidity score could be calculated. This information may be used as an important decision support tool to guide therapy selection, to realistically estimate the need for resources, and to plan the care for high-risk patients more efficiently. By monitoring outcomes and resource utilizations using computerized patient records, the performance of the decision process may be continuously evaluated.

Chapter 7

Conclusions

The major conclusions reached were:

I. The additive EuroSCORE algorithm had a significantly better discriminatory power to predict 30-day mortality than the STS risk algorithm in CABG-only surgery.

II. The EuroSCORE, Cleveland Clinic and Magovern risk algorithms showed superior performance and accuracy in open-heart surgery, and the EuroSCORE, NYS and Cleveland Clinic scoring systems in CABG-only surgery. Though originally designed to predict early mortality, the one-year mortality prediction was also reasonably accurate.

III. The additive EuroSCORE algorithm can be used to predict total cost, ICU cost and an ICU stay of more than two days after open-heart surgery.

IV. By using an ANN risk stratification model, risk factors in a ranked order contributing to the operative mortality prediction could be identified. A minimal set of risk variables achieving a superior mortality prediction could be defined. The ANN model was applicable independent of the cardiac surgical procedure.

Acknowledgements

This thesis has been accomplished thanks to the help and support of many people to whom I owe my gratitude. In particular, I would like to express my sincere appreciation to:

Associate Professor Johan Brandt, my supervisor and friend, for invaluable scientific guidance, for continuous support and encouragement, and for many fruitful and interesting discussions.

Associate Professor Peter Höglund, statistical expert and co-author, for sharing his extensive knowledge and experience in statistics, for many stimulating discussions, continuous encouragement and friendship.

Dr. Lars Algotsson, my co-supervisor, for sharing his knowledge and experience in anaesthesia and intensive care, and for his continuous encouragement.

Mattias Ohlsson, expert and my teacher in ANN methodology, for sharing his knowledge and experience in machine learning techniques and make them understandable.

Dr. Carsten Lührs, my clinical mentor, co-author and friend, for his patience and support when I started my training in cardiothoracic surgery, and for sharing his database knowledge, which made this thesis possible.

My co-authors Lars Thulin and Samer Nashef, founders of the EuroSCORE database, who made *Study IV* possible to perform.

Professor Stig Steen, head of cardiothoracic research, for introducing me to the field of science and for fruitful scientific discussions.

Colleagues and friends at the Department of Cardiothoracic Surgery and the Department of Cardiothoracic Anaesthesia and Intensive Care, for their encouragement and friendship, and for making this thesis possible by entering the patient data in the database.

Our secretary, Birgitta Sjögren, for helping me with the patient records and for always being supportive.

Economist Camilla Andersson, for helping me with the cost calculations in *Study III*.

My parents-in-law, Birgitta and Gunnar, for always supporting me and my family and lending me the "scientific cottage" in Åhus where I finished my thesis.

My parents, Kerstin and Gösta, for all their support and love, and for showing me the way in life.

To the two most beautiful and lovely girls in the world, my twins, Sofie and Hanna. Their gorgeous wake-up smile every morning gave daddy energy to finish this work.

And finally my dear Bodil, my soulmate and the love of my life, for being my discussion partner any time of the day or place in the world, for always supporting me, and for encouraging me to keep on and finish this thesis.

The research presented in this thesis has been supported by grants from The Swedish Heart and Lung Foundation, Anna Lisa and Sven-Eric Lundgrens foundation for medical research and by computer resources from Lunarc at Lund University.

References

1. Nightingale F. Sanitary Condition of Hospitals. *Notes on Hospital.* 3 ed. London: Longman, Green, Longman, Roberts and Green; 1863:1-6.

2. Codman E. A study in Hospital Efficiency as Demonstrated by the Case Report of the First Five Years of a Private Hospital. Boston: Thomas Todd Company; 1917.

3. Cochrane A. Effectiveness & Efficiency. *Random Reflections on Health Service*. London: Royal Society of Medicine Press Ltd; 1972.

4. Brinkley J. U.S. releasing list of hospitals with abnormal mortality rates. *New York Times.* March 12, 1986: 1.

5. Kouchoukos NT, Ebert PA, Grover FL, Lindesmith GG. Report of the ad hoc committee on risk factors for coronary artery bypass surgery. *Ann Thorac Surg.* Mar 1988;45(3):348-349.

6. O'Connor GT, Plume SK, Olmstead EM, Coffin LH, Morton JR, Maloney CT, et al. A regional prospective study of in-hospital mortality associated with coronary artery bypass grafting. The Northern New England Cardiovascular Disease Study Group. *JAMA.* Aug 14 1991;266(6):803-809.

7. Hannan EL, Kilburn H, Jr., Racz M, Shields E, Chassin MR. Improving the outcomes of coronary artery bypass surgery in New York State. *JAMA.* Mar 9 1994;271(10):761-766.

8. Hammermeister KE, Johnson R, Marshall G, Grover FL. Continuous assessment and improvement in quality of care. A model from the Department of Veterans Affairs Cardiac Surgery. *Ann Surg.* Mar 1994;219(3):281-290.

9. 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.* Feb 1 1995;91(3):677-684.

10. Bridgewater B, Neve H, Moat N, Hooper T, Jones M. Predicting operative risk for coronary artery surgery in the United Kingdom: a comparison of various risk prediction algorithms. *Heart.* Apr 1998;79(4):350-355.

11. Edwards FH, Clark RE, Schwartz M. Coronary artery bypass grafting: the Society of Thoracic Surgeons National Database experience. *Ann Thorac Surg.* Jan 1994;57(1):12-19.

12. Grover FL, Hammermeister KE, Shroyer AL. Quality initiatives and the power of the database: what they are and how they run. *Ann Thorac Surg.* Nov 1995;60(5):1514-1521.

13. Bernstein AD, Parsonnet V. Bedside estimation of risk as an aid for decision-making in cardiac surgery. *Ann Thorac Surg.* Mar 2000;69(3):823-828.

14. Hannan EL, Kilburn H, Jr., O'Donnell JF, Lukacik G, Shields EP. Adult open heart surgery in New York State. An analysis of risk factors and hospital mortality rates. *JAMA.* Dec 5 1990;264(21):2768-2774.

15. Chassin MR. Improving quality of care with practice guidelines. *Front Health Serv Manage.* Fall 1993;10(1):40-44.

16. Andersson B, Nilsson J, Brandt J, Höglund P, Andersson R. Gastrointestinal complications after cardiac surgery. *Br J Surg.* Mar 2005;92(3):326-333.

17. Sjögren J, Nilsson J, Gustafsson R, Malmsjö M, Ingemansson R. The impact of vacuum-assisted closure on long-term survival after post-sternotomy mediastinitis. *Ann Thorac Surg.* Oct 2005;80(4):1270-1275.

18. Ferraris VA, Ferraris SP. Risk Stratification and Comorbidity. In: Cohn LH, Edmonds LHJ, eds. *Cardiac surgery in the adult*. 2 ed. New York: McGraw-Hill Companies; 2003:187-224.

19. 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*. Jun 1989;79(6 Pt 2):I3-12.

20. Knaus WA, Wagner DP, Zimmerman JE, Draper EA. Variations in mortality and length of stay in intensive care units. *Ann Intern Med*. May 15 1993;118(10):753-761.

21. Coronary Artery Bypass Surgery in New York State 1996-1998. *New York State Department of Health* [pdf]. Available at: http://www.health.state.ny.us/nysdoh/consumer/heart/1996-98cabg.pdf. Accessed April 8, 2005.

22. Grover FL, Johnson RR, Marshall G, Hammermeister KE. Factors predictive of operative mortality among coronary artery bypass subsets. *Ann Thorac Surg*. Dec 1993;56(6):1296-1306; discussion 1306-1297.

23. Grover FL, Shroyer AL, Hammermeister KE. Calculating risk and outcome: the Veterans Affairs database. *Ann Thorac Surg*. Nov 1996;62(5 Suppl):S6-11; discussion S31-12.

24. 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*. Sep 28 1999;100(13):1464-1480.

25. O'Connor GT, Plume SK, Olmstead EM, Coffin LH, Morton JR, Maloney CT, et al. Multivariate prediction of in-hospital mortality associated with coronary artery bypass graft surgery. Northern New England Cardiovascular Disease Study Group. *Circulation*. Jun 1992;85(6):2110-2118.

26. 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. *JAMA.* May 6 1992;267(17):2344-2348.

27. Jones RH, Hannan EL, Hammermeister KE, Delong ER, O'Connor GT, 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.* Nov 15 1996;28(6):1478-1487.

28. Ferraris VA, Ferraris SP, Singh A. Operative outcome and hospital cost. *J Thorac Cardiovasc Surg.* Mar 1998;115(3):593-602; discussion 602-593.

29. Riordan CJ, Engoren M, Zacharias A, Schwann TA, Parenteau GL, Durham SJ, et al. Resource utilization in coronary artery bypass operation: does surgical risk predict cost? *Ann Thorac Surg.* Apr 2000;69(4):1092-1097.

30. Pintor PP, Bobbio M, Colangelo S, Veglia F, Marras R, Diena M. Can EuroSCORE predict direct costs of cardiac surgery? *Eur J Cardiothorac Surg.* Apr 2003;23(4):595-598.

31. Sokolovic E, Schmidlin D, Schmid ER, Turina M, Ruef C, Schwenkglenks M, et al. Determinants of costs and resource utilization associated with open heart surgery. *Eur Heart J.* Apr 2002;23(7):574-578.

32. Pintor PP, Colangelo S, Bobbio M. Evolution of case-mix in heart surgery: from mortality risk to complication risk. *Eur J Cardiothorac Surg.* Dec 2002;22(6):927-933.

33. Osswald BR, Tochtermann U, Schweiger P, Gohring D, Thomas G, Vahl CF, et al. Minimal early mortality in CABG-simply a question of surgical quality? *Thorac Cardiovasc Surg.* Oct 2002;50(5):276-280.

34. Osswald BR, Blackstone EH, Tochtermann U, Thomas G, Vahl CF, Hagl S. The meaning of early mortality after CABG. *Eur J Cardiothorac Surg.* Apr 1999;15(4):401-407.

35. Daley J. Criteria by which to evaluate risk-adjusted outcomes programs in cardiac surgery. *Ann Thorac Surg.* Dec 1994;58(6):1827-1835.

36. Edwards FH, Clark RE, Schwartz M. Practical considerations in the management of large multiinstitutional databases. *Ann Thorac Surg.* Dec 1994;58(6):1841-1844.

37. Iezzoni LI. The risks of risk adjustment. *JAMA.* Nov 19 1997;278(19):1600-1607.

38. Hannan EL, Racz MJ, Jollis JG, Peterson ED. Using Medicare claims data to assess provider quality for CABG surgery: does it work well enough? *Health Serv Res.* Feb 1997;31(6):659-678.

39. Lee T. *Evaluating the Quality of Cardiovascular Care: A Primer*: American College of Cardiology Press; 1995.

40. Jamieson WR, Edwards FH, Schwartz M, Bero JW, Clark RE, Grover FL. Risk stratification for cardiac valve replacement. National Cardiac Surgery Database. Database Committee of The Society of Thoracic Surgeons. *Ann Thorac Surg.* Apr 1999;67(4):943-951.

41. Altman DG. Categorising continuous variables. *Br J Cancer.* Nov 1991;64(5):975.

42. Schafer JL. Multiple imputation: a primer. *Stat Methods Med Res.* Mar 1999;8(1):3-15.

43. Little RJA. Regression with missing X's: A review. *Journal of the American Statistical Association.* 1992;87:1227-1237.

44. Steen PM. Approaches to predictive modeling. *Ann Thorac Surg.* Dec 1994;58(6):1836-1840.

45. 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.* Jun 2004;77(6):2232-2237.

46. Hosmer DW, Lemeshow S. Applied logistic regression. New York: John Wiley & Sons; 2000.

47. Harrell FE, Jr., Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med.* Feb 28 1996;15(4):361-387.

48. Dupont W. *Statistical modeling for biomedical researchers*. Cambridge: Cambridge University Press; 2002.

49. Schemper M, Smith TL. Efficient evaluation of treatment effects in the presence of missing covariate values. *Stat Med.* Jul 1990;9(7):777-784.

50. Kirkwood BR, Sterne JAC. Bayesian statistics. *Essential Medical Statistics*. 2 ed. Malden: Blackwell Science Ltd; 2003:388-390.

51. Schemper M, Heinze G. Probability imputation revisited for prognostic factor studies. *Stat Med.* Jan 15-Feb 15 1997;16(1-3):73-80.

52. Warren S. How to measure importance of inputs? *SAS Institute Inc.* June 23, 2000. Available at: ftp://ftp.sas.com/pub/neural/importance.html. Accessed June 23, 2005.

53. Bishop CM. *Neural Networks for Pattern Recognition*. Oxford: Oxford University Press; 2003.

54. Wehrens R, Putter H, Buydens L. The bootstrap: a tutorial. *Chemom Intel Lab Syst.* 2000;54:35-52.

55. Eferon B, Tibshirani R. *An introduction to the bootstrap*. New York: Chapman & Hall; 1993.

56. Marshall G, Shroyer AL, Grover FL, Hammermeister KE. Bayesian-logit model for risk assessment in coronary artery bypass grafting. *Ann Thorac Surg.* Jun 1994;57(6):1492-1499; discussion 1500.

57. 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.* Jun 2003;75(6):1856-1864; discussion 1864-1855.

58. Shahian DM, Blackstone EH, Edwards FH, Grover FL, Grunkemeier GL, Naftel DC, et al. Cardiac surgery risk models: a position article. *Ann Thorac Surg.* Nov 2004;78(5):1868-1877.

59. Michel P, Roques F, Nashef SA. Logistic or additive EuroSCORE for high-risk patients? *Eur J Cardiothorac Surg.* May 2003;23(5):684-687.

60. Nashef SA, Roques F, Michel P, Gauducheau E, Lemeshow S, Salamon R. European system for cardiac operative risk evaluation (EuroSCORE). *Eur J Cardiothorac Surg.* Jul 1999;16(1):9-13.

61. Shwartz M, Ash A, Iezzoni L. Comparing outcomes across providers. In: Iezzoni L, ed. *Risk Adjustment for Measuring Healthcare Outcomes*. Chicago: Health Administration Press; 1997:471.

62. Thomas N, Longford NT, Rolph JE. Empirical Bayes methods for estimating hospital-specific mortality rates. *Stat Med.* May 15 1994;13(9):889-903.

63. 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.* Dec 2001;72(6):2155-2168.

64. Goldstein H, Spiegelhalter D. League tables and their limitations: statistical issues in comparisons of institutional performance. *J R Stat Soc.* 1996;159:385.

65. Kirkwood BR, Sterne JAC. Random effects (multilevel) models. *Essential Medical Statistics*. 2 ed. Malden: Blackwell Science Ltd; 2003:361-366.

66. Carey JS, Robertson JM, Misbach GA, Fisher AL. Relationship of hospital volume to outcome in cardiac surgery programs in California. *Am Surg.* Jan 2003;69(1):63-68.

67. Zelen J, Bilfinger TV, Anagnostopoulos CE. Coronary artery bypass grafting. The relationship of surgical volume, hospital location, and outcome. *N Y State J Med.* Jul 1991;91(7):290-292.

68. Haykin S. *Neural networks. A compressive foundation*. 2 ed. Upper Saddle River: Prentice Hall; 1999.

69. 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.* Apr-Jun 1998;18(2):229-235.

70. Lippmann RP, Shahian DM. Coronary artery bypass risk prediction using neural networks. *Ann Thorac Surg.* Jun 1997;63(6):1635-1643.

71. 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.* Feb 15 1997;79(4):857-862.

72. DeGroff CG, Bhatikar S, Hertzberg J, Shandas R, Valdes-Cruz L, Mahajan RL. Artificial neural network-based method of screening heart murmurs in children. *Circulation.* Jun 5 2001;103(22):2711-2716.

73. Heden B, Öhlin H, Rittner R, Edenbrandt L. Acute myocardial infarction detected in the 12-lead ECG by artificial neural networks. *Circulation.* Sep 16 1997;96(6):1798-1802.

74. Baxt WG. Application of artificial neural networks to clinical medicine. *Lancet.* Oct 28 1995;346(8983):1135-1138.

75. Orr RK. Use of a probabilistic neural network to estimate the risk of mortality after cardiac surgery. *Med Decis Making.* Apr-Jun 1997;17(2):178-185.

76. Ennett CM, Frize M. Weight-elimination neural networks applied to coronary surgery mortality prediction. *IEEE Trans Inf Technol Biomed.* Jun 2003;7(2):86-92.

77. Buzatu DA, Taylor KK, Peret DC, Darsey JA, Lang NP. The determination of cardiac surgical risk using artificial neural networks. *J Surg Res.* Jan 2001;95(1):61-66.

78. 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. *AMIA Annu Symp Proc.* 2003:160-164.

79. Gunn S. *Support Vector Machines for Classification and Regression.* Southampton: Image Speech & Intelligent Systems Group; 1998.

80. Harrell FE, Jr. *Regression modeling strategies with applications to linear models, logistic regression, and survival analysis.* New York: Springer-Verlag; 2001.

81. Anderson RP, Jin R, Grunkemeier GL. Understanding logistic regression analysis in clinical reports: an introduction. *Ann Thorac Surg.* Mar 2003;75(3):753-757.

82. Shwartz M, Ash A. Evaluating the performance of risk-adjustment methods: continuous measures. In: Iezzoni L, ed. *Risk Adjustment for Measuring Health Care Outcomes.* Chicago: Health Administration Press; 1994:287.

83. Lusted LB. Logical analysis in roentgen diagnosis. *Radiology.* Feb 1960;74:178-193.

84. Swets JA. Measuring the accuracy of diagnostic systems. *Science.* Jun 3 1988;240(4857):1285-1293.

85. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology.* Apr 1982;143(1):29-36.

86. Bamber D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *J Math Psychol.* 1975;12:387-415.

87. Altman DG. *Practical statistics for medical research.* London: Chapman & Hall; 1999.

88. Kirkwood BR, Sterne JAC. Assessing reproducibility of measurements. *Essential Medical Statistics.* 2 ed. Malden: Blackwell Science Ltd; 2003:434.

89. Asimakopoulos G, Al-Ruzzeh S, Ambler G, Omar RZ, Punjabi P, Amrani M, et al. An evaluation of existing risk stratification models as a tool for comparison of surgical performances for coronary artery bypass grafting between institutions. *Eur J Cardiothorac Surg.* Jun 2003;23(6):935-942.

90. Geissler HJ, Holzl P, Marohl S, Kuhn-Regnier F, Mehlhorn U, Sudkamp M, et al. Risk stratification in heart surgery: comparison of six score systems. *Eur J Cardiothorac Surg.* Apr 2000;17(4):400-406.

91. 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.* Feb 2002;21(2):199-204.

92. Rexius H, Brandrup-Wognsen G, Nilsson J, Oden A, Jeppsson A. A simple score to assess mortality risk in patients waiting for coronary artery bypass grafting. *Ann Thorac Surg (In press).* 2005.

93. Huijskes RVHP, Rosseel PMJ, Tijssen JGP. Outcome prediction in coronary artery bypass grafting and valve surgery in the Netherlands: development of the Amphiascore and its comparison with the Euroscore. *European Journal of Cardio-Thoracic Surgery.* 2003;24(5):741-749.

94. Kurki TS, Kataja M. Preoperative prediction of postoperative morbidity in coronary artery bypass grafting. *Ann Thorac Surg.* Jun 1996;61(6):1740-1745.

95. Roques F, Michel P, Goldstone AR, Nashef SA. The logistic EuroSCORE. *Eur Heart J.* May 2003;24(9):882.

96. Roques F, Gabrielle F, Michel P, De Vincentiis C, David M, Baudet E. Quality of care in adult heart surgery: proposal for a self-assessment approach based on a French multicenter study. *Eur J Cardiothorac Surg.* 1995;9(8):433-439; discussion 439-440.

97. 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.* Nov 1 1996;28(5):1147-1153.

98. Gabrielle F, Roques F, Michel P, Bernard A, de Vicentis C, Roques X, et al. Is the Parsonnet's score a good predictive score of mortality in adult cardiac surgery: assessment by a French multicentre study. *Eur J Cardiothorac Surg.* Mar 1997;11(3):406-414.

99. Pons JM, Granados A, Espinas JA, Borras JM, Martin I, Moreno V. Assessing open heart surgery mortality in Catalonia (Spain) through a predictive risk model. *Eur J Cardiothorac Surg.* Mar 1997;11(3):415-423.

100. Ivanov J, Tu JV, Naylor CD. Ready-made, recalibrated, or remodeled? Issues in the use of risk indexes for assessing mortality after coronary artery bypass graft surgery. *Circulation.* Apr 27 1999;99(16):2098-2104.

101. Ivanov J, Borger MA, David TE, Cohen G, Walton N, Naylor CD. Predictive accuracy study: comparing a statistical model to clinicians' estimates of outcomes after coronary bypass surgery. *Ann Thorac Surg.* Jul 2000;70(1):162-168.

102. Tremblay NA, Hardy JF, Perrault J, Carrier M. A simple classification of the risk in cardiac surgery: the first decade. *Can J Anaesth.* Feb 1993;40(2):103-111.

103. 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.* Jul 1992;102(1):36-44.

104. 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.* Jun 2004;77(6):1966-1977.

105. Eagle KA, Guyton RA. ACC/AHA 2004 guideline update for coronary artery bypass graft surgery [pdf]. Available at: http://www.acc.org/clinical/guidelines/cabg/index.pdf. Accessed October 15, 2005.

106. Parsonnet V. Risk stratification in cardiac surgery: is it worthwhile? *J Card Surg.* Nov 1995;10(6):690-698.

107. Jones RH. In search of the optimal surgical mortality. *Circulation.* Jun 1989;79(6 Pt 2):I132-136.

108. Ferraris VA. The dangers of gathering data. *J Thorac Cardiovasc Surg.* Jul 1992;104(1):212-213.

109. Green J, Wintfeld N. Report cards on cardiac surgeons. Assessing New York State's approach. *N Engl J Med.* May 4 1995;332(18):1229-1232.

110. 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.* Jun 1999;15(6):816-822; discussion 822-813.

111. Hjortso E, Buch T, Ryding J, Lundstrom K, Bartram P, Dragsted L, et al. The nursing care recording system. A preliminary study of a system for assessment of nursing care demands in the ICU. *Acta Anaesthesiol Scand.* Oct 1992;36(7):610-614.

112. Kirkwood BR, Sterne JAC. Comparing two proportions. *Essential Medical Statistics.* 2 ed. Malden: Blackwell Science Ltd; 2003:148-164.

113. Hosmer DW, Lemeshow S. Model-building strategies and methods for logistic regression. In: Hosmer DW, Lemeshow S, eds. *Applied logistic regression.* 2 ed. New York: John Wiley & Sons; 2000:91-134.

114. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics.* Sep 1988;44(3):837-845.

115. Cross SS, Harrison RF, Kennedy RL. Introduction to neural networks. *Lancet.* Oct 21 1995;346(8982):1075-1079.

116. Pepe MS. The receiver operating characteristic curve. *The Statistical Evaluation of Medical Tests for Classification and Prediction.* New York: Oxford University Press; 2003:92-94.

117. Roques F, Nashef SA, Michel P, Pinna Pintor P, David M, Baudet E, et al. Does EuroSCORE work in individual European countries? *Eur J Cardiothorac Surg.* Jul 2000;18(1):27-30.

118. Nashef SA, Roques F, Hammill BG, Peterson ED, Michel P, Grover FL, et al. Validation of European system for cardiac operative risk evaluation (EuroSCORE) in North American cardiac surgery. *Eur J Cardiothorac Surg.* Jul 2002;22(1):101-105.

119. Pliam MB, Shaw RE, Zapolanski A. Comparative analysis of coronary surgery risk stratification models. *J Invasive Cardiol.* Apr 1997;9(3):203-222.

120. MaWhinney S, Brown ER, Malcolm J, VillaNueva C, Groves BM, Quaife RA, et al. Identification of risk factors for increased cost, charges, and length of stay for cardiac patients. *The Annals of Thoracic Surgery.* 2000;70(3):702-710.

121. Smith PK, Smith LR, Muhlbaier LH. Risk stratification for adverse economic outcomes in cardiac surgery. *Ann Thorac Surg.* Dec 1997;64(6 Suppl):S61-63; discussion S80-62.

122. Kurki TS, Hakkinen U, Lauharanta J, Ramo J, Leijala M. Evaluation of the relationship between preoperative risk scores, postoperative and total length of stays and hospital costs in coronary bypass surgery. *Eur J Cardiothorac Surg.* Dec 2001;20(6):1183-1187.

123. Kurki TS, Kataja MJ, Reich DL. Validation of a preoperative risk index as a predictor of perioperative morbidity and hospital costs in coronary artery bypass graft surgery. *Journal of Cardiothoracic and Vascular Anesthesia.* 2002;16(4):401-404.

124. Lawrence DR, Valencia O, Smith EE, Murday A, Treasure T. Parsonnet score is a good predictor of the duration of intensive care unit stay following cardiac surgery. *Heart.* Apr 2000;83(4):429-432.

125. Stoica SC, Sharples LD, Ahmed I, Roques F, Large SR, Nashef SA. Preoperative risk prediction and intraoperative events in cardiac surgery. *European Journal of Cardio-Thoracic Surgery.* 2002;21(1):41-46.

126. Pitkanen O, Niskanen M, Rehnberg S, Hippelainen M, Hynynen M. Intra-institutional prediction of outcome after cardiac surgery: comparison between a locally derived model and the EuroSCORE. *Eur J Cardiothorac Surg.* Dec 2000;18(6):703-710.

127. 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(3):258-264.

128. Tu JV, Guerriere MR. Use of a neural network as a predictive instrument for length of stay in the intensive care unit following cardiac surgery. *Comput Biomed Res.* Jun 1993;26(3):220-229.

129. Rognvaldsson T, You L. Why neural networks should not be used for HIV-1 protease cleavage site prediction. *Bioinformatics.* Jul 22 2004;20(11):1702-1709.

130. Williams TE, Jr., Fanning WJ, Benton WC, Kakos GS, Miller RL, Esterline WJ, et al. What is the marginal cost for marginal risk in cardiac surgery? *Ann Thorac Surg.* Dec 1998;66(6):1969-1971.

131. Michel P, Domecq S, Rachid Salmi L, Roques F, Nashef SA. Confidence intervals for the prediction of mortality in the logistic EuroSCORE (reply). *Eur J Cardiothorac Surg.* Jun 2005;27(6):1129-1132.

132. Murphy-Filkins R, Teres D, Lemeshow S, Hosmer DW. Effect of changing patient mix on the performance of an intensive care unit severity-of-illness model: how to distinguish a general from a specialty intensive care unit. *Crit Care Med.* Dec 1996;24(12):1968-1973.

Papers I-IV

Johan Nilsson

Risk Stratification in Cardiac Surgery

Lund 2005

FACULTY OF MEDICINE
Lund University

Lund University, Faculty of Medicine Doctoral Dissertation Series 2005:102

ISSN 1652-8220
ISBN 91-85481-03-3