INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

Bell & Howell Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

UMI®

Université d'Ottawa · University of Ottawa

CORONARY SURGERY MORTALITY PREDICTION USING ARTIFICIAL NEURAL NETWORKS

by

Colleen Michelle Ennett,

B.Sc. (Eng), Biological Engineering, University of Guelph

A thesis submitted to the
School of Graduate Studies and Research
in partial fulfilment of the requirements for the degree of

Master of Applied Science
in Electrical Engineering

Ottawa-Carleton Institute for Electrical and Computer Engineering
School of Information Technology and Engineering

Faculty of Engineering
University of Ottawa

November 1999

In loving memory of my dad, Mike Ennett,
who taught me that learning is a lifelong process,

and

to my mom, Gaye (Heartwell) Ennett,
and my brother, Craig Ennett,
for their constant support and encouragement.

Abstract

This thesis demonstrates the application of a feedforward backpropagation-trained artificial neural network using the weight-elimination cost function to the estimation of in-hospital mortality for coronary artery bypass grafting patients from the San Francisco Heart Institute in Daly City, California, USA. The highly-skewed *a priori* statistics due to the low mortality rate present difficulties for modelling this data. Artificial training and test datasets with higher mortality rates were developed to improve the classification performance of the artificial neural networks. Sensitivity was considered the most important measure of performance for this work. Given that current mortality risk models are unable to accurately identify high-risk patients (those who do not survive the surgery), focussing on increasing the sensitivity rate will indicate when more of the patients who are difficult to classify are correctly identified. The final result was an increase in sensitivity when training with a dataset with a higher mortality rate than the test set. This dataset modification approach resulted in only small changes for other performance measures (specificity, predictive positive value, predictive negative value, and correct classification rate), and thus helped to achieve the main goal of the study.

Acknowledgements

First and foremost, I would like to acknowledge the support of my thesis supervisor, Dr. Monique Frize, who guided me through the past two years. Thank you for believing in me from the start. Dr. Frize has been a wonderful supervisor, mentor, and friend to me. She always offered constructive advice, and her support encouraged me to strive higher. I am grateful for her financial assistance throughout my master's degree funded through her Medical Research Council grant CGAA-45088. Looking forward to working on my doctorate under your supervision.

I thank Dr. Shirley Mills for her statistical insight to ensure that my approaches were precise and accurate. I appreciated the assistance that Dr. Richard E. Shaw offered whenever I needed an expert medical opinion. Thank you to the research group at San Francisco Heart Institute for providing the cardiac surgery patient database used in this thesis research. Many thanks to Yanling Tong and the others in our research group, MIRG (Medical IDEAS Research Group), for their input on my thesis.

I am grateful to my family for providing moral support. And a special thanks to Tariq Haddad for offering his thesis writing expertise and for stimulating my interest in research.

Contents

List of Figures

List of Tables

Nomenclature

α	momentum
λ	weight-elimination constant
ANN	artificial neural network
ASE	average squared error
CABG	coronary artery bypass graft
CAD	coronary artery disease
CCR	correct classification rate
COPD	chronic obstructive pulmonary disease
CP	constant predictor
CPM	conditional probability matrix
CVA	cerebrovascular accident
CVOR	Consortium for Virtual Operations Research
DECH	Doctor Everett Chalmers Hospital
err_ratio	error ratio
FN	false negative
FP	false positive
GCS	Glasgow coma score
IABP	intraaortic balloon pump
ICU	intensive care unit
IV	intravenous
lr	learning rate
lr_inc	learning rate adaptive parameter to increase learning rate
lr_dec	learning rate adaptive parameter to decrease learning rate
MI	myocardial infarction
MIRG	Medical IDEAS (Intelligent Decision Aid Systems) Research Group
MLP	multilayer perceptron
PNV	predictive negative value
PPV	predictive positive value
PTCA	percutaneous transluminal coronary angioplasty
ROC	receiver operating characteristic curves
RNG	random number generator
SFHI	San Francisco Heart Institute
SSE	sum of squared error
STS	Society of Thoracic Surgeons
teorig	test set with the true in-hospital mortality rate (3.8 percent)
te20	artificial test set with an in-hospital mortality rate of 20 percent
TN	true negative
TP	true positive
trorig	training set with the true in-hospital mortality rate (3.7 percent)
tr10	artificial training set with an in-hospital mortality rate of 10 percent

tr20	artificial training set with an in-hospital mortality rate of 20 percent
tr30	artificial training set with an in-hospital mortality rate of 30 percent
w_0	weight-elimination scale factor
WE	weight-elimination
wtfact	output error weighting factor

"In attempting to arrive at the truth, I have applied everywhere for information, but in scarcely an instance have I been able to obtain hospital records fit for any comparison. If they could be obtained, they would enable us to decide many other questions than the one alluded to."

— Florence Nightingale, Notes on Hospitals, 1873

Chapter 1: Introduction

The arrival of the information age spawned a renewed interest in the prediction of future outcomes. There is an unspoken belief that a "magic combination" of information exists that will make reliable predictions. This search for the "magic combination" has led to an accumulation of data on specific events. For example, some medical researchers believe that if doctors could identify the higher risk patients and the factors causing this increased risk, precautions could possibly be taken to reduce or eliminate this threat. This type of risk stratification has been investigated concerning coronary artery bypass grafting (CABG) surgery.

The search for an effective method of mortality risk stratification for coronary artery surgery began in 1986 after the Health Care Financing Administration in the United States began releasing raw statistics on the mortality rate of Medicare coronary artery bypass surgery patients in American hospitals. The stated objective was to inform patients about the quality of care at hospitals, and to help them make knowledgeable decisions to attain the best service and treatment possible [Ebert 1989]. This mortality data, from the point of view of the hospitals and surgeons, did not consider the patient's severity of illness before undergoing surgery.

Categorizing the patients into different levels of risk provides a more accurate view of the quality of surgical care, and can potentially be used as a decision aid to assess a patient's risk of mortality before surgery. Presently, these risk assessment models are not sufficiently accurate at the individual level to permit use in a clinical setting. The models are used to observe changes in the characteristics of the patient population over a period of years, effects of changes in surgical, pre- and postoperative procedures, and statistical variations from institution to institution. The heart surgery patient population is also a particularly difficult group to classify. It appears that there are few defining characteristics that easily identify whether a patient will survive the surgery or not [Orr 1997].

As the processing power and storage capacity of today's computers grow, so does the need for an effective method of interpreting this data. Medical researchers have tried various methods of risk stratification using additive models, statistical approaches, and a form of artificial intelligence called artificial neural networks (ANNs). Additive models are intuitive, since understanding how having more risk factors increases the risk of death is easy. Statistical techniques are commonly relied upon to derive patterns in large amounts of data to predict outcomes for new patients. ANNs, a newer approach whose popularity resurged in the late 1980's, can use nonlinear modelling algorithms to find the most suitable model to fit the data. Schematically and functionally based on their biological counterparts, ANNs have artificial neurons and variable weights to simulate the actions of the synapse to permit learning. This form of artificial intelligence has achieved success in other medical settings [Baxt 1991, 1994a, 1994b, Buskard 1994, Buskard *et al.* 1994, Frize *et al.* 1995, 1996, 1997, Trigg 1997], so improvements in outcome predictions for CABG surgery patients are anticipated.

1.1 Motivation and Significance of the Research

In the past decade, many risk models have been developed and updated to estimate the risk of mortality for coronary artery surgery patients. To date, however, no model has gained widespread acceptance due to poor generalizability to institutions other than the originating one. Most researchers have found that the best models are institution-specific. In a discussion about the search for the best coronary artery risk model, W.C. Nugent [1995] clearly described the situation by saying:

> "Rather than arguing whether one model is preferable to another, the central question is whether a validated mathematical model is preferable to no model at all in helping to make clinical decisions. The answer to this is irrefutably yes."

With the medical advances that have increased the mean age of our society, a new population is demanding coronary artery surgery. These individuals are older, with more health problems, have more repeat surgeries, poorer ventricular function, and more diffuse disease [Keon & Menzies 1992]. Even with advances in surgical techniques and myocardial protection, surgeons still have difficulties combatting the increased risk of operative death and morbidity for this new, frail population. Despite this, nearly all medical institutions discussed in the literature report an open-heart surgery mortality rate of less than 5 percent.

Usually, coronary artery surgery mortality models can easily make accurate predictions about low-risk patients, while the higher-risk patients are poorly stratified. The challenge presented by the CABG database is the identification of these patients who are at a high risk of death. Although a low mortality rate is desirable from a medical point of view, the small number of nonsurvivors presents a serious challenge for risk prediction techniques. Having too few samples poses several problems. Most notably for an ANN, this means there is not likely enough information to learn which combination of factors will lead to death. This problem of a small representative sample of one outcome is not unique to coronary artery surgery. In many other fields, prediction of an underrepresented class is desirable, but so far no foolproof solution exists. If a solution to the small amount of information available to researchers about coronary artery surgery mortality rates could be found, the approach could possibly be suitable in other domains as well.

1.2 Thesis Objective

This study's approach will use a feedforward artificial neural network (ANN) trained with the backpropagation learning rule and the weight-elimination cost function to predict the in-hospital mortality of CABG surgery patients at the San Francisco Heart Institute (SFHI) in Daly City, California. The effect of changing the distribution of the training and test datasets by entering the cases of the underrepresented class multiple times into the artificial dataset will be examined. Artificially changing the *a priori* statistics of the datasets is an attempt to combat the problems that arise with the low representation of the nonsurvivors of CABG surgery when predicting in-hospital mortality using ANNs.

As a second objective, the largest network weights at the point of optimal performance of the ANN will be extracted to identify the variables that have the most influence on the predictions of the outcome. This goal will be accomplished using the weight-elimination error cost function. These variables will be extracted from the database, and the performance of the ANN with only these "most important" variables will be evaluated.

The application of this ANN algorithm with weight-elimination to a medical database has given successful results in a previous study [Trigg 1997, Frize *et al.* 1997]. In her thesis, Trigg's problem of identifying intensive care unit (ICU) patients requiring 8 or more hours of mechanical ventilation involved an approximately 70-30 percent distribution of the outcomes. This is a much larger percentage

of the underrepresented category than the problem presented by the CABG database. For CABG surgery outcomes, the patients who do not survive the operation generally represent less than 5 percent of the database. This poses serious difficulties for the intended ANN to be investigated, as shown by preliminary results done by Trigg [1997] and in a conference paper by Ennett and Frize [1998]. Ennett and Frize [1998] discovered a region of inconsistent network performance (i.e., sometimes the network would learn patterns and other times it would classify everything as belonging to the group with the highest *a priori* probability) when the underrepresented class made up approximately 15 percent of the cases. At representations of less than 15 percent, the ANN failed to learn anything. Instead, it classified all patients as belonging to the larger outcome class (i.e., those patients requiring less than 8 hours of artificial ventilation). A copy of the full paper by Ennett and Frize [1998] is included in Appendix A.[1] In theory, this technique of changing the *a priori* statistics of the training set may provide a solution to the challenges faced when dealing with a drastically underrepresented outcome in a two-class problem.

1.3 Thesis Outline

This section delineates the chapters briefly to identify the topics covered in each section of this thesis. In Chapter 2, an introduction to the basic concepts covered in this thesis is provided. There is a brief description of the medical environment being investigated, followed by a description of the analysis techniques that are commonly used for evaluating risk models. Furthermore, a review of the earlier work by the research groups ensues, and the various techniques that other researchers have employed to separate the CABG patients into categories of increasing risk are mentioned according to the type of model involved: additive models, statistical techniques, and artificial intelligence.

Chapter 3 outlines the problem formulation. First, a discussion of the challenges involved with CABG surgery databases that one must keep in mind arises, and a list of possible solutions for these difficulties is presented. The next part discusses advancements that have occurred in ANN research, followed by an overview of the SFHI database including statistical descriptions of the information contained in the database, and a justification of the choice of the reduced dataset. Then, the techniques that can be employed for identifying the most important variables are discussed.

[1] Used with permission, see Appendix B.

Chapter 4 justifies the ANN architecture that will be employed in developing the mortality prediction model, and the techniques used to create the training and test sets. Various sampling techniques are described for creating the artificial datasets. There is a short discussion about the choice of experiments selected to obtain useful and concise results.

In Chapter 5, the general results of the experimental simulations and their evaluation are presented. A description of the network parameter optimization ensues. The results of each particular set of experiments are analyzed separately to extract the important information that was found. The thesis results are compared with previous work by Trigg [1997] and the SFHI research group [Pliam *et al.* 1997] to validate the findings. Finally the section concludes with some general comments about the experimental results.

Chapter 6 summarizes the final results of the work completed, and identifies the areas where new information was uncovered. Finally, suggestions for future work ideas in this area of medical research and artificial neural networks complete this thesis.

Chapter 2: An Overview of Relevant Concepts

For a more thorough understanding of the medical aspects of the thesis topic, some background information about the medical environment involved with this research is provided. A description of the techniques that researchers use to analyse the performance of the existing coronary artery risk stratification models provides the groundwork for comparing different models. A review of the previous work accomplished by the research groups involved, the Medical IDEAS Research Group (MIRG, IDEAS = Intelligent Decision Aid Systems) and the San Francisco Heart Institute (SFHI) research group, serves as a backdrop for this thesis work. Before beginning a discussion of the methodology, knowing what techniques other researchers have tried is important, as well as how well these approaches have succeeded in stratifying the patients into the appropriate risk categories. There are three main types of models for risk stratification: additive models, statistical techniques, and artificial neural networks. Within each category of models, descriptions of several examples that are commonly used as benchmarks by which to compare newer models are provided.

2.1 Medical Context

The poor diet and sedentary lifestyle of a typical North American have allowed the incidence of coronary artery disease (CAD) to flourish and become Canada's number one killer. These are not the only factors that can increase a person's chance of developing this disease, however, they are key players in the disease's onset. CAD can begin as angina pectoris which occurs when the coronary arteries become constricted or blocked preventing an adequate supply of blood to the heart muscle. Without sufficient oxygen, myocardial infarction (MI) or sudden death could result [Youngson 1992].

Initially, medications such as nitrates, beta-blockers, and calcium inhibitors can treat patients with angina pectoris [Maurice & Lancelin 1980]. Medical treatment alleviates the pain associated with angina, and although it reduces the chance of occurrence of MI and sudden death, patients are still at risk. As the disease progresses to involve multiple cardiac blood vessels, or the left ventricular function becomes impaired, percutaneous transluminal coronary angioplasty (PTCA) and/or coronary artery bypass graft surgery (CABG) may be necessary. An angioplasty offers a less invasive approach to treating CAD. It has many of the same benefits of CABG, but is not quite as effective. CABG can offer substantial or complete pain relief, and the risks of MI and sudden death are drastically reduced. These improvements allow the patient to enjoy a better quality of life. At the same time, however, surgery introduces new risks such as operative death, graft occlusion, aortic dissection and stroke [Miller 1977].

A patient with heart disease may undergo isolated CABG surgery or a combined CABG surgery. "Isolated" surgery means that the only purpose of the operation is to perform the CABG procedure. A "combined" surgery simply indicates that more than one procedure is being performed. This means that, in addition to the CABG operation, the surgeons will also perform another procedure such as valve replacement or repair, or an endarterectomy, among other procedures. All surgical procedures have an inherent risk of morbidity and mortality (caused by unforeseen complications).

Recall that the mortality rate for CABG surgery is close to 5 percent. This is a low mortality rate, and shows that surgeons are generally successful at recognizing suitable candidates for coronary artery surgery. A morbidity outcome such as a stroke or an infection resulting from the surgical procedure can sometimes be corrected using medical expertise, but death cannot be reversed. Knowing this, there is no question that preoperatively identifying whether a patient will survive surgery is critical.

2.2 Classifier Evaluation Techniques

To assess the validity of a risk stratification model, the approach must be evaluated using commonly accepted measures of performance. Below is an overview of several measures that are generally accepted as effective and relatively unbiased. The typically uneven *a priori* probabilities of the medical output variables adds to the challenge of developing an effective model [Penny & Frost 1996]. Given the highly skewed data from the SFHI, finding a measure that is not highly influenced by

the low mortality rate is difficult. It is hoped, however, that using several validation techniques will provide a better overall picture of the model's performance.

2.2 (a) Constant Predictor

A constant predictor is a basic statistical benchmark that classifies all cases as belonging to the class with the greatest *a priori* probability [Frize *et al.* 1995]. To illustrate this, consider the case of CABG surgery. The outcome of interest is in-hospital death. The percentage of patients in the training set who die is 3.7 percent (meaning that 96.3 percent survive the surgery), and in the test set, 3.8 percent of patients die. A constant predictor would classify all patients in the test set as surviving, and it would, therefore, correctly classify the patients 96.2 percent of the time.

2.2 (b) Contingency Table

Table 2.1 describes a truth table. The correct classification rate (CCR), or accuracy, identifies the rate at which the model correctly classifies the data into their proper categories. The CCR is calculated by summing the number of cases that were correctly classified into their respective classes (the number of true positives, TP, plus the number of true negatives, TN, and dividing by the total number of cases in the dataset (TP+TN)/total cases). The sensitivity of the model tells the rate at which a subject who dies following surgery is classified as dying. The formula for sensitivity is TP/(FN + TP) where FN represents the number of false negatives. Specificity is the rate at which a patient who survives is classified as surviving, and can be calculated using the following formula: TN/(TN + FP) where FP is the number of false positives. Both the sensitivity and specificity are properties of the model's test. The predictive positive value (PPV) is the rate at which the cases are classified as dying following surgery, TP/(FP + TP), and the predictive negative value (PNV) is the rate at which the cases are classified as surviving the surgery, TN/(TN + FN). These parameters depend on the prevalence of the situation under investigation [Penny & Frost 1996]. Although these may be effective means of measuring performance in some cases, the prevalence of the outcomes can be influential.

Table 2.1: Contingency table

		Correct Output		
		Not Present	**Present**	
Model Output	**Not Present**	true negative (TN)	false negative (FN)	TN + FN
	Present	false positive (FP)	true positive (TP)	FP + TP
		TN + FP	FN + TP	total cases

An important point to consider is the cost of misclassification. There may be different costs associated with the misclassification of a patient. For example, predicting that a patient will not survive surgery (but the patient actually lives) has a different associated cost, than foretelling survival, when in actual fact the patient will die. In this situation of different misclassification costs, the CCR is not necessarily the best measure of accuracy [Kattan & Beck 1995]. Although this is the situation with CABG data, the CCR will be reported knowing the limitations of its interpretation. Also, a greater focus will be on the sensitivity rate, since this measure is more sensitive to the correct classification of the nonsurvivors of coronary artery surgery who are more difficult to classify. Finding a satisfactory balance between the sensitivity, specificity and CCR would be ideal.

2.2 (c) Mean Predicted Mortality

This measure is calculated by simply summing the raw predicted mortality values for each case, and dividing by the number of cases. The mean predicted mortality is the average predicted risk of all patients in the group. It can offer an overview of the type of patients in the test set such as whether there were a large number of higher-risk patients (indicated by a higher mean predicted mortality), and it can give an idea of how well the model is calibrated to the test data by comparing the mean predicted mortality with the actual mortality rate.

2.2 (d) Receiver Operating Characteristic Curves

The area under the receiver operating characteristic curves (ROC curves, or the C-index) assesses the ability of the model to discriminate between outcomes. Since this measure does not require a predefined decision threshold, it may also be used to discover the optimal cutpoint for the test [Forsström & Dalton 1995]. It is a plot of the model's sensitivity versus one minus its specificity. In medical terminology, the area under the ROC curve is the probability of a positive test given no disease versus the probability of a positive test given the disease. The generated ROC curve is a visual description of the operating points and potential tradeoff between the true and false positive rates. This curve is obtained by varying the threshold value of the output node across its range of values [Woods & Bowyer 1997]. Despite being a well-accepted measure of performance, the C-index can still be influenced by the prevalence of the outcome of interest [Buchman *et al.* 1994].

ROC curve analysis only works for two output class problems since the generated curve is two-dimensional. Figure 2.1 shows several sample ROC curves. A perfect model would be represented by an ROC curve that is a step function. This would indicate that all values of the true positive rate are equal

to one (i.e., no classification error). The ROC curve of a useless classifier is a 45-degree positive diagonal line, where the true positive rate equals the false positive rate, and offers no improvement over random guessing. Thus, the closer the ROC curve resembles a step function, the better the model is. Typical C-index values for coronary artery surgery models range from 0.70 to 0.80.

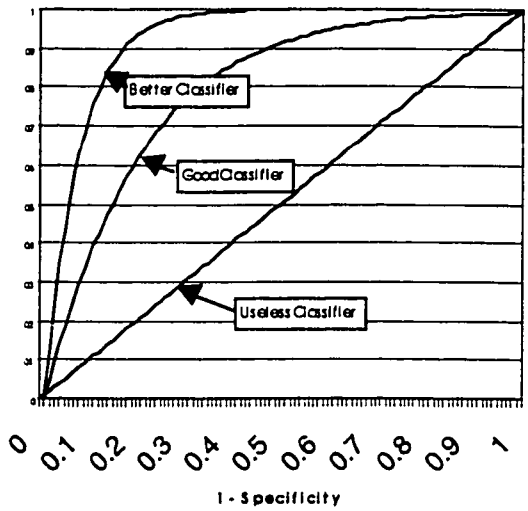


Figure 2.1: Sample ROC Curves

2.3 The Medical IDEAS Research Group

MIRG is a multidisciplinary research group that draws on the skills and knowledge of the engineers, computer scientists, statisticians, physicians, and students involved. MIRG's objective is to use artificial intelligence to integrate several decision-aid systems in a way that fits well with the way physicians work. Applied to medicine, artificial intelligence should simulate common clinical thinking to aid and support decision-making and patient management processes. Among MIRG's ongoing projects is the application of ANNs to the estimation of clinical outcomes. The ANN simulates the clinician's approach: "And for this particular patient, this is what I think will happen" [Frize *et al.* 1995]. MIRG has explored various medical domains, including adult and neonatal ICUs, chronic arthritis management, and childhood cancer.

MIRG research is guided by its principal investigators (university professors and clinicians), and progress is made through the completion graduate level theses and term projects, in addition to

undergraduate senior theses. This section briefly outlines the work completed by former master's level graduate students.

2.3 (a) Tim Buskard's Work

Using a database of 1322 medical and surgical adult intensive care unit (ICU) patients from the Doctor Everett Chalmers Hospital (DECH) in Fredericton, New Brunswick, the preliminary work with ANNs by Tim Buskard involved the retrospective estimation of duration of artificial ventilation, mortality, and length of stay [Buskard 1994, Buskard *et al.* 1994]. Mortality and length of stay are common outcomes to investigate in an ICU setting, however, the prediction of the duration of artificial ventilation appears to have been a novel outcome of interest.

Based on his literature search to find the most effective ANN, Buskard chose to use a feedforward backpropagation ANN. Buskard attempted to classify patients into 10 output classes for the variable "length of stay" (0, 1, 2, 3, 4, 5, 6, 7, 8-14 and >14 days), and 12 output classes for "duration of artificial ventilation" (0, 0-12, <12-24, <24-48, <48-72, ..., <336-504 and >504 hours) using 41 input parameters. The ANN was trained on two-thirds of the cases, and tested on the remaining one-third of the database. With so few training examples and so many output classes, the ANN was vulnerable to overfitting (or memorization). The overfitting to the training set was evidenced by the divergence of the CCR curves of the training and test sets after only a few hundred epochs. Although the ANN achieved higher CCRs than a constant predictor, the improvements were marginal. These results could be attributed to the complexity of the model.

Buskard's work showed that a feedforward backpropagation ANN is an effective classification tool for the medical environment. He also discovered that a poor balance of the number of inputs, output classes and training data (i.e., the complexity of the model) leads to overfitting to the training set, and poor generalization to the test set.

2.3 (b) Heather Trigg's Work

Heather Trigg (nee McGowan) continued the work of Buskard by attempting to eradicate the problem of overfitting. She used two new approaches to classify the data: weight-elimination and high-low node representation [Trigg 1997, Frize *et al.* 1997]. Her objective was to improve the CCR without losing potentially useful information from the database or waiting for enough additional patient cases to be collected. The weight-elimination approach (described in more detail in Section 2.5 where the

basics of ANNs are introduced) was supposed to prevent overfitting by removing irrelevant weights. The goal of the high-low node data presentation technique was to see if isolating whether a parameter was higher-than-normal, normal, or lower-than-normal would improve the classification accuracy. This idea was based on the fact that several physiological parameters have different health effects depending on these characteristics [Frize *et al.* 1996, 1997]. Trigg used the same database as Buskard with a few additional cases (1491 total), and the cases were separated into medical and surgical ICU patients. Since the surgical (postoperative) patients have more common physiological characteristics than medical (nonpostoperative) patients, Trigg chose to present the postoperative patient cases (883) to the ANN.

To focus on the performance of the ANN for the two situations under consideration, only two outcomes were investigated: less than 8 hours of artificial ventilation, and greater than or equal to 8 hours of artificial ventilation. Trigg used a total of 51 input variables for the regular data presentation, and 65 inputs for the high-low node technique. Unfortunately, Trigg discovered that the high-low node representation did not improve the CCR of the ANN for this particular database, however, the double-layered ANNs slightly improved the performance for this data. The optimal network performance occurred using a double-layered ANN with weight-elimination (CCR = 91.8 percent). Trigg's configuration achieved a statistically significant improvement over the minimum distance classifier (CCR = 86.1 percent) and the constant predictor (CCR = 71.1 percent). The combination of the high-low node representation and weight-elimination techniques with a single-layered network allowed the most important variables for predicting artificial ventilation of more than eight hours to be extracted from the network [Trigg 1997, Frize *et al.* 1997, 1998]. The most important parameters were defined as those with the largest absolute values after weight elimination.

A double-layered network offered a slight improvement in the CCR for the ICU database. Trigg's most interesting discovery, however, was the extraction of the largest weights from the ICU database remaining in the single-layered ANN after weight-elimination. Theoretically, weight-elimination will cause the weights of the least influential variables to drop to zero, ultimately falling out of the model's equation. This finding is important because it suggests the potential for the ANN to select the significant variables for outcome prediction without bias from the researcher.

2.4 The San Francisco Heart Institute Research Group

Pliam and his colleagues, a group of researchers at the SFHI, compared existing coronary artery surgery risk models against each other using their own database. They developed a logistic regression formula from the results, and assessed their predictive accuracy with ROC curves [Pliam *et al.* 1997]. When the SFHI research group analysed the current methods of predicting mortality for CABG patients, they looked at additive models (the Parsonnet model and the Cleveland Clinic model), and statistical models (Bayesian theory and logistic regression). Two of the risk models using Bayesian theory (STS1 and STS2) were taken from the literature. These models were developed on two different databases, one being a database from a single hospital and the other from the Society of Thoracic Surgeons (STS) Adult Cardiac National Database. Pliam *et al.* [1997] also developed Bayesian and logistic regression models based on the SFHI database. The observed mortality rate for the SFHI patients with combined CABG procedures (CABG only, CABG plus valve surgery, CABG plus repair surgery) was 4.0 percent in the test set, while that of the isolated CABG patients was 3.7 percent.

A summary of the variables used in each risk model investigated by the SFHI Research Group is found in Table 2.2. Some models were developed using specific patient populations, such as isolated CABG surgery, and Table 2.3 describes which models are intended for which population.

Table 2.2: Risk factors in additive and statistical CABG surgery risk studies

Risk Factor	Parsonnet Model	Cleveland Clinic	Edwards *et al.* Bayes'	SFHI Bayes'	SFHI Log Reg	STS Version 1	STS Version 2
Age	X	X	X	X	X	X	X
Female gender	X		X	X		X	X
Date of surgery							X
Valve (aortic/mitral) surgery	X	X			X		
CABG plus valve surgery	X						
Emergency/urgent surgery	X	X	X	X	X	X	X
Repeat operation	X	X	X	X	X	X	X
Prior vascular surgery		X					
Renal failure/dysfunction			X	X			X
Dialysis dependency	X						
Elevated creatinine level		X					
Left ventricular function	X	X	X			X	
Left ventricular aneurysm	X		X			X	
Left main disease			X	X	X	X	X
Low ejection fraction				X		X	X
Mitral insufficiency		X					
Aortic stenosis	X	X		X			X
Coronary dissection						X	
Valvular disease			X	X	X		X
One-, 2-, & 3-vessel disease			X	X		X	X
Pulmonary hypertension	X						
Hypertension	X		X	X		X	
Preoperative intraaortic balloon pump (IABP)	X			X	X	X	
Previous myocardial infarction (MI)			X	X	X	X	
Acute evolving MI				X		X	
MI-When							X
Congestive heart failure				X			
Unstable angina			X				X
Cerebrovascular disease		X	X	X			
Peripheral vascular disease				X	X	X	
Morbid obesity	X		X	X			X
Weight 65 kg/small stature		X		X			
Chronic obstructive pulmonary disease (COPD)		X	X	X			X
Diabetes	X	X	X	X			X
Cardiogenic shock			X	X	X	X	X
Current tobacco abuse			X	X	X		
> 100 pack-years smoking			X	X			
Nonsmoker				X			
Intravenous (IV) nitrates			X			X	
IV inotropic support			X			X	
Cardiopulmonary resuscitation						X	
PTCA emergency			X	X			X
Prior/failed PTCA					X	X	X
Hypercholesterolemia				X			
Previous cerebrovascular accident (CVA)/CVA				X	X		X
Cardiomegaly							X
Anemia		X		X			
Catastrophic states	X						
Rare circumstances	X						
Total number of variables	17	14	23	29	13	20	21

Table 2.3: Datasets used in additive and statistical coronary artery surgery risk studies

Data Set	Parsonnet Model	Cleveland Clinic	Edwards *et al.* (Bayes')	SFHI (Bayes')	SFHI (Log Reg)	STS Version 1	STS Version 2
Isolated CABG surgery		X	X			X	X
CABG plus valve surgery		X					
All open-heart surgeries	X			X	X		

The results of the SFHI research group showed that separating the patients with isolated CABG surgery from those with combined procedures did not improve the model's performance. Including patients undergoing combined procedures increases the range of patients that can benefit from this model. Pliam *et al.* [1997] concluded that these CABG models could accurately classify outcomes about 80 percent of the time. Sometimes differences in risk model performance can be attributed to the differences in the particular variables selected as risk factors [Nugent 1995]. The SFHI database did not always have the same variables as required for the models from literature, but Pliam *et al.* [1997] detailed how they dealt with the missing data by either following the instructions outlined in the articles of how to treat the missing data, substituting with a variable which offers similar information, or omitting the variable completely. Table 2.4 provides a summary of their results using the SFHI database [Pliam *et al.* 1997].

Table 2.4: Summary of SFHI results using various risk stratification models

Risk Model	Mean Predicted Mortality	C-index
Parsonnet	9.0 ± 8.0	0.80 ± 0.02
Cleveland Clinic	6.0 ± 6.0	0.80 ± 0.02
SFHI Bayesian	7.6 ± 15.6	0.83 ± 0.02
SFHI logistic regression	5.1 ± 7.7	0.80 ± 0.02
Parsonnet (isolated CABG)	8.4 ± 7.4	0.80 ± 0.03
Cleveland Clinic (isolated CABG)	5.7 ± 5.9	0.80 ± 0.03
SFHI Bayesian (isolated CABG)	6.5 ± 13.9	0.83 ± 0.02
SFHI logistic regression (isolated CABG)	4.5 ± 6.5	0.79 ± 0.03
STS Version 1 (isolated CABG)	9.6 ± 9.1	0.77 ± 0.03
STS Version 2 (isolated CABG)	3.0 ± 3.3	0.81 ± 0.02

The findings of Pliam and his colleagues [1997] highlight that it is not necessary to separate isolated CABG patients from those with combined procedures when using additive or statistical models for risk stratification of coronary artery surgery patients. This is an interesting discovery since it is often believed that a more homogeneous database will improve the model's performance.

2.5 CABG Surgery Risk Stratification Approaches

A literature search revealed several modelling strategies applied to the problem of assessing risk of death following CABG surgery. The models described in the subsequent sections represent only a sampling of the numerous coronary artery surgery risk models available in the literature. These models include additive models, statistical methods, and ANNs. Examples of the most commonly used models in each category are also described briefly.

2.5 (a) Additive Models

Additive models are sometimes preferred to other modelling techniques because they are more intuitive for clinicians (the cumulative effect of more risk factors can easily be associated with added risk), and can be computed using only a basic calculator. Typically, additive models do not perform well on higher-risk patients. Their risk of death may be considerably underestimated which is unacceptable in practice because the intention of risk models is an easier identification of these higher-risk patients [Clark 1996]. Despite this limitation, the two additive models compared by Pliam *et al.* [1997] will be briefly discussed. The Parsonnet and Cleveland Clinic models are among the earlier CABG surgery mortality risk models developed, and have been used as benchmarks by which to compare the performance of newer techniques.

2.5 (a) i. Parsonnet's Model

The univariate additive model developed by Parsonnet and his colleagues [1989] is among the most referred to coronary artery surgery mortality models in the literature and many institutions have attempted to implement it. Several research institutions based their initial data collections on the risk factors identified by the Parsonnet group [Higgins *et al.* 1992, Edwards *et al.* 1994a, Pliam *et al.* 1997, etc.]. The main goal of this model was to stratify open-heart surgery patients into mortality risk categories using readily available data for input variables. Parsonnet *et al.* [1989] defined mortality as death within 30 days of surgery. It is important to note that this additive model was designed to predict operative mortality of open-heart surgery patients, that is, patients undergoing any type of heart surgery including CABG.

The Parsonnet model was derived using a retrospective analysis of 3500 consecutive open-heart surgery procedures at the Newark Beth Israel Medical Center spanning five years (1982-1987). Univariate regression analysis identified 17 significant mortality risk factors, as outlined in Table 2.2.

A variable had to meet certain inclusion criteria before being accepted as a risk factor. The inclusion criteria had to demonstrate predictive value by univariate analysis, universal availability for every patient and at every institution, freedom from subjectivity or bias, and the variable must be simple and direct (i.e., not derived from other information) [Parsonnet *et al.* 1989].

Parsonnet *et al.* [1989] chose to weigh the risk factors based on the value of the odds ratio that compares the chance of mortality with the specific risk factor. The additive model was developed in the form $p=k_1+k_2+...+k_{17}$, where p is the predicted probability of operative mortality and k_1 to k_{17} are the non-zero risk scores assigned to each applicable risk factor. If the risk factor is not present in the patient's case, then it is excluded from the calculation. The model classified the patients as belonging to one of five groups of increasing risk: good (0-4%), fair (5-9%), poor (10-14%), high (15-19%), and extremely high (≥20%). The mean predicted operative mortality was 10.4 percent and the mean observed operative mortality was 8.9 percent for the test set of 1332 patients from the originating institution [Parsonnet *et al.* 1989]. There was no mention of the mortality rate for the training set. This mortality rate is notably higher than those reported at most other institutions (where the mortality rates are usually less than 5 percent).

The additive Parsonnet model was tested at two other clinical institutions with satisfactory results [Parsonnet *et al.* 1989]. Although Parsonnet's model attempts to be objective, two of the risk factors ("catastrophic states" and "other rare circumstances") rely on the individual physician's opinion to assign weights to these factors [Higgins *et al.* 1992]. These subjective risk factors hinder the model's generalizability because these variables are ill-defined and may be interpreted differently by different surgeons.

2.5 (a) ii. Cleveland Risk Model

Higgins *et al.* [1992] from the Cleveland Clinic Foundation developed an additive model referred to as the Cleveland Risk Model. Again, this is another model that has been a benchmark for CABG surgery model performance. The Cleveland group felt that, although mortality was an important outcome for investigation, morbidity was perhaps a better indication of the quality of care provided by the hospital. Based on this observation, their additive model aimed to separate CABG patients into different levels of morbidity *and* mortality risk based on their preoperative severity of illness using data that spanned 1986 to 1990. The model only identified the risk of morbidity in general, not the specific morbidity for which the patient was at risk. The outcomes that Higgins *et al.* [1992] considered as

morbid were myocardial infarction, use of intra-aortic balloon pump (IABP), mechanical ventilation for three or more days, neurological deficit, oliguric or anuric renal failure, and serious infection [Higgins *et al.* 1992]. This summary will discuss the mortality risk model. The research group defined mortality as death during the hospitalization for surgery, regardless of length of stay, or within 30 days of hospital discharge. An additional goal when selecting risk factors for the additive model was that the items be routinely collected prior to surgery to avoid the need for special testing, and that the factors be relatively free of physician judgement.

CABG operations accompanied by additional surgical procedures, such as carotid endarterectomy and heart valve replacements, were included within the definition of "CABG surgery." This is simply because the population undergoing these procedures, in addition to CABG, represents an important portion of the surgical patients in actual practice (according to Higgins *et al.* [1992]).

Using univariate and logistic regression analysis on the 5051 patient cases in the training set, the significant risk factors related to perioperative mortality were identified as listed in Table 2.2. The univariate odds ratios, the degree of significance in the logistic model, and clinical considerations lead to the selection of the weight for each risk factor. The value of the weight for each factor was varied to optimize the performance of the additive model. A panel of several physicians involved in the study provided input for the selection of the final weight for each risk factor. Weight values for the risk factors range from 1 to 6 points to predict mortality and morbidity. Theoretically, this simple additive score has a maximum severity score of 31 points.

The model was developed based on a mortality rate of 2.5 percent, but 3.6 percent of patients in the test group died postoperatively (n = 4169). The patient risk scores were categorized into nine different categories of increasing risk (scores of 0, 1, 2, 3, 4, 5, 6, 7-9, $\geq$10). The differences in the distribution of the high-risk patients lead to a greater occurrence of higher-risk patient scores (score $\geq$ 5) in the test group.

2.5 (b) Statistical Models

The framework of statistics is the calculation of probabilities. This discipline is regulated by theorems, demonstrations, assumptions, and hypothesis testing [Vicino 1998]. Although statistical models are the most commonly used technique, they are less intuitive than additive models. The mathematical equations for Bayes' Theorem and logistic regression models are complicated to calculate

by hand, and in a clinical situation would require a programmable calculator or a computer program. Despite the added complexity (compared with additive models), medical researchers have relied heavily on statistical techniques to model this type of medical data.

2.5 (b) i. Bayes' Theorem

This statistical technique was made popular among medical researchers by Edwards *et al.* [1987] who provided the basic principles in an article where he simplified this approach. The theorem of Bayes uses information theory principles, combining the effects of several factors affecting a single patient. The main advantage of Bayes' theorem is its ability to predict outcomes for an individual patient. The *a priori* probabilities for a training group of patients are applied to Equation 2.1. The absence or presence of the particular risk factors results in the calculation of a risk probability for the patient. This technique uses the distribution of risk factors among a reference group to predict the probability of a future event. In the case of predicting the risk of mortality associated with coronary artery surgery, operative death is the outcome probability of interest, and the *a priori* statistics of the most significant risk factors are used as the input information. As with most other statistical models, the patient risk factors must first be identified using regression analysis to observe their relationship with mortality.

The predictive risk equation of Bayes' theorem is:

$$P_i = \frac{\prod_{j=1}^{J} \{a_j P(S_j|D_i) + (1-a_j)[1- P(S_j|D_i)]\} P(D_i) f(P)}{\sum_{i=1}^{I} \prod_{j=1}^{J} \{a_j P(S_j|D_i) + (1-a_j)[1- P(S_j|D_i)]\} P(D_i)} \qquad (2.1)$$

where P_i is the probability of outcome I, D_i corresponds to one of I possible outcomes, S_j corresponds to one of the J risk factors, $P(S_j|D_i)$ is the conditional probability that risk factor S_j is present in outcome D_i, and $f(P)$ is an exponential correction factor. The toggle parameter $a_j = 1$ when the associated risk factor S_j is present; otherwise $a_j = 0$. The prior probabilities are denoted by $P(D_i)$ [Edwards *et al.* 1987]. For the coronary surgery mortality risk model, P_i would be the probability of an operative death. This probability is calculated by dividing the number of deaths by the total number of patients. In the simplest terms, the conditional probabilities are the fraction of patients with outcome D_i who have risk factor S_j present [Edwards *et al.* 1987].

Bayes' theorem uses a conditional probability matrix (CPM) to relate the rate of occurrence of a particular event with the desired outcome in the training set. These probabilities in the CPM become the constants of the Bayesian equation (Equation 2.1), and the presence or absence of these events are the equation variables. Once these probabilities in the training database are calculated and put into the Bayesian equation, the model is tested against a new population (test set) to observe its predictive capability.

The major drawback of Bayes' theorem, however, is that it only provides a snapshot of the frequency of the risk factors at the time of the training set data collection. The Bayesian equation assumes that the population is unchanged over time, since it does not take into consideration changes in treatment patterns or surgical techniques. To keep the equation up-to-date, the CPM should be updated once every few years, or when substantial changes are made to procedures.

2.5 (b) ii. Logistic Regression Model

The most popular method of regression analysis used by medical researchers is logistic regression [Kleinbaum *et al.* 1998]. Its popularity is based on its ability to model dichotomous dependent variables, as well as to produce an output that varies between 0 and 1 that resembles a probability estimate. This is particularly useful in the medical field because researchers would often like to provide a probability of risk for a certain person to have a particular disease, for example. Logistic regression is structurally equivalent to a feedforward ANN with linear inputs and one output unit with a sigmoidal nonlinear transfer function [Lippmann *et al.* 1995].

The logistic model is described in Equation 2.2 as the probability of occurrence of one of the two possible outcomes [Kleinbaum *et al.* 1998]. This regression model can describe the association between several predictor variables $X_1, X_2, \ldots, X_k$ and a dichotomous dependent variable Y, where Y is usually coded as 1 or 0 for its two possible outcome categories. The parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ are the regression coefficients that need to be estimated using the training dataset.

$$pr(Y = 1) = \frac{1}{1 + \exp\left[-\left(\beta_0 + \sum_{j=1}^{k} \beta_j X_j\right)\right]} \tag{2.2}$$

2.5 (b) iii. Society of Thoracic Surgeons' Risk Models

Edwards *et al.* [1987] continued their research by developing models using the Bayesian theorem as outlined in Section 2.5(b)i. [Edwards *et al.* 1988, 1994a], and the method of logistic regression described in Section 2.5(b)ii. [Shroyer *et al.* 1998, 1999]. Data collection in the United States for the STS Adult Cardiac National Database began in 1980. Practice groups, hospitals, and surgeons participate on a voluntary basis. The STS cardiac database contained approximately 1.2 million patient records in 1997, and approximately 500 sites submit data to the STS database [Shroyer *et al.* 1999]. The STS models by Edwards *et al.* [1988, 1994a] and Shroyer *et al.* [1998, 1999] are based on isolated CABG surgery patients to maintain a more homogenous database. Edwards and his colleagues saw the opportunity to develop a risk stratification model using a national database. Collecting information from institutions across the country may be advantageous, because they believed they could then develop a generalized model that would be universally applicable.

The first model using Bayes' Theorem, STS1, was based on isolated CABG surgery patient data at one institution. The 700 cases of patient data for STS1 were collected between January 1984 and April 1987. They were arranged in the database chronologically, and the first 300 patients comprised the training set. The remaining 400 patients were sequentially evaluated in groups of 100 patients, after which the group was incorporated into the training set, and the model was updated. The objective was to observe the model's performance as it was revised with time. The 20 risk factors were chosen based on clinical intuition, and mortality was defined as death within 30 days of CABG. The mortality rate of the test group was 4.75 percent. Here, the patients were sorted into categories of risk: <5, 5-25, 25-50, >50, >80, >90 percent. Edwards *et al.* [1988] did not use any statistical techniques to evaluate their models. Instead, they compared the predicted versus observed mortality rates for the risk categories, and concluded that all risk groups agreed [Edwards *et al.* 1988].

In the early years of the STS Database (1980-83), only a few institutions were collecting data. Moreover, the mid-1980's saw a well-recognized change in the CABG population, so Edwards *et al.* [1994a] chose to only use records of isolated CABG from 1984 to 1990 to develop the second Bayesian model (STS2). For the STS2 Bayesian model, the risk factors for isolated CABG surgery were selected based on the results of univariate analysis and multivariate analysis using stepwise logistic regression. The mortality rate for the entire population was 3.2 percent (n = 78,927), and the Bayesian model grouped the patients into five mortality risk categories (0.0-2.05, 2.05-3.9, 3.9-6.75, 6.75-11.90, >11.90

percent) [Edwards *et al.* 1994a]. The parameters deemed important for both STS1 and STS2 are shown in Table 2.2.

Research continued with the STS National Database, and two logistic regression isolated CABG models were developed: the 1995 Coronary Artery Bypass Risk Model [Shroyer *et al.* 1998] and the 1996 Coronary Artery Bypass Risk Model [Shroyer *et al.* 1999]. These models used data from 1990-95 and 1990-96, respectively. The training set for the 1995 model had 69 458 patient cases with a 3.1 percent mortality rate, whereas the mortality rate for the 87 271 patients in the 1996 model was 3.2 percent. The most noticeable change was the inclusion of racial background as a risk factor. Between the two new models, New York Heart Association class IV, use of steroids, use of digitalis, and serum creatinine level were risk factors that the STS had not previously considered. These factors had been used in risk models developed at other institutions. In each model, patient risk was separated into one of seven risk categories (0-2.5, 2.5-5.0, 5-10, 10-20, 20-30, 30-50, 50-100 percent) [Shroyer *et al.* 1998, 1999].

As discussed previously, the main drawback to using the STS database is the fact that it is a large database. Sometimes important relationships can be washed out in large databases, or other insignificant relationships can be inflated so that they appear to be more influential than they actually are. Prudence is the key when working with large sets of data. Also, most hospitals participating in STS database collection are large institutions, so the dynamics of smaller hospitals are not as obvious [Ebert 1989]. Edwards *et al.* [1994b] recognized this problem, but believed that the bias would have less influence once more smaller practices were incorporated into the STS database.

2.5 (c) Artificial Neural Networks

Artificial neural networks (ANNs) are a form of artificial intelligence — "computers that learn." The main inspiration for the development of ANNs was likely the desire to mimic human thought processes with an artificial system. Ideally, the artificial system could perform sophisticated, even "intelligent," computations in a similar manner to the routine functioning of the human brain. The development of ANNs also may be an attempt to shed light on the inner workings of the human brain.

ANNs are "parallel, distributed, adaptive information-processing systems that develop their functionality in response to exposure to information" [Penny & Frost 1996]. The popularity of ANNs applied to medicine began in the early 1990's following their reintroduction into the realm of prediction

techniques in the mid 1980's by Rumelhart *et al.* [1986a, 1986b]. The advantage that ANNs offer over statistical techniques is that the model does not have to be explicitly defined before the experiments begin. ANNs can grasp the relevant data to develop the model, whereas to derive a statistical model, prior knowledge of the relationships between the factors under investigation is required [Blum 1992, Livingstone *et al.* 1997].

Recall that the structure and function of ANNs are based on the interconnected network of the brain. Like a biological neural network, the ANN has simple processing units called artificial neurons, or nodes, as illustrated in Figure 2.2. The node has inputs where the known information about the situation is introduced. These correspond to the receiving ends of the synapse of a biological neuron. The inputs are then each multiplied by a corresponding weight, or "synaptic connection." The values are then summed in each processing unit, and fed through a transfer function to scale the output to a value within a limited range (i.e., between 0 and 1, or between -1 and 1). Each neuron has a bias input that is always equal to one. The bias input also has a weight associated with it to determine its impact on the network. The bias input is multiplied by its weight (resulting in a value that is not necessarily 1), and its function is comparable to a nonzero intercept in statistics [Kattan & Beck 1995]. The bias weight provides a threshold above which the node is activated [Penny & Frost 1996]. The node's output becomes the input of the next layer, or if already at the output layer, the resulting signals are the network output [Itchhaporia *et al.* 1996].

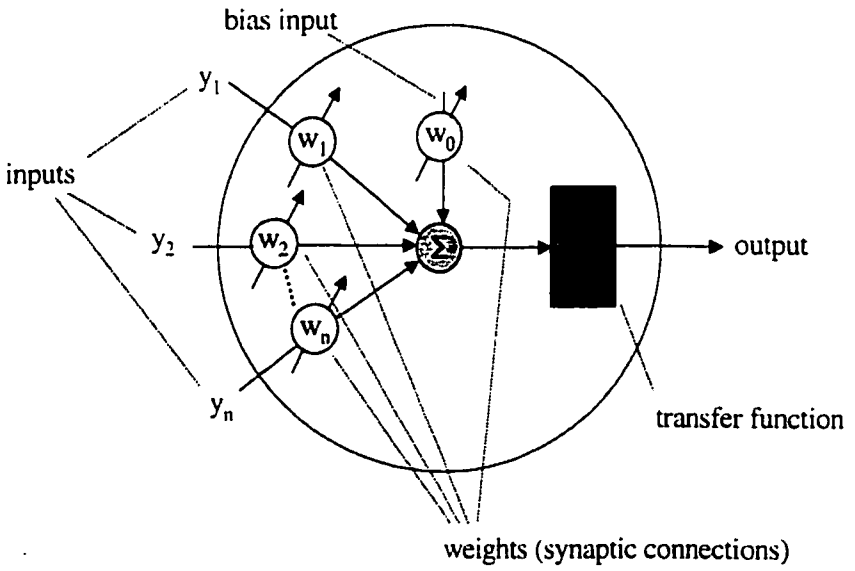


Figure 2.2: Diagrammatic representation of an artificial neuron

The number of artificial neurons in a hidden layer greatly influences the performance of the network. If there are too few nodes, there will not be sufficient resources to solve the problem, and underfitting results. On the other hand, too many nodes may cause long training times, and possibly memorization rather than generalization ("overfitting") [Masters 1993]. Despite this, the number of nodes and inputs should not be arbitrarily reduced simply to avoid overfitting, because underfitting may occur. Hence, a method of network pruning should be employed to reduce the network size.

The nodes in a network are connected in parallel to create layers as shown in Figure 2.3. As seen in the diagram, a node is connected to every node in the layer below it. A network can be described in a short form that indicates the number of inputs, nodes in the hidden layer(s), and outputs. For example, the ANN in Figure 2.3 would be described as a 4-3-1 network, because it has four inputs, three nodes in the hidden layer, and one output node.

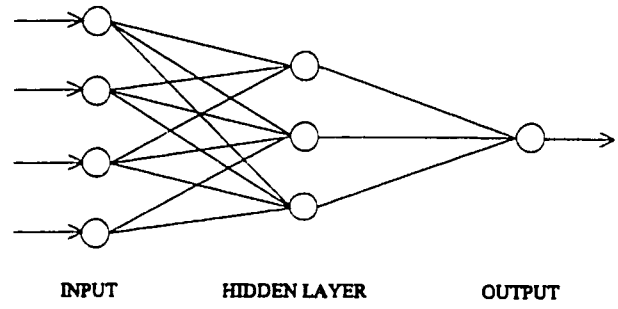


Figure 2.3: Architecture of a simple feedforward artificial neural network (4-3-1 network)

In a feedforward ANN, otherwise known as a multilayer perceptron (MLP), the information travels forward from the inputs to the outputs. If any of the outputs are fed back as inputs to previous layers, the network is called a "feedback" or recurrent ANN [Fausett 1994, Jang *et al.* 1997]. Typically, when reporting the number of layers in a network the input layer is not included because the raw data is entered at this level and no data manipulation occurs. Depending on the database, the network may or may not require hidden layers. Hidden layers allow more complicated problems to be solved, however, training is more difficult because there are more parameters to optimize. Finally, the output layer reveals the values of the interlayer weights that relate inputs to outputs for the given database [Itchhaporia *et al.* 1996].

An ANN with no hidden layers is a single-layered ANN, and is similar to the generalized linear model from statistics. A network with one hidden layer (thus, a double-layered ANN), however, can adequately describe almost any function to any desired degree of accuracy, given that the function and its derivative are continuous, and the ANN uses sufficient hidden units [Hornik *et al.* 1989, Fausett 1994, Penny & Frost 1996]. As of yet, no one has proved with substantial evidence that more than one hidden layer improves the predictive performance of a network [Penny & Frost 1996].

The information from the input signal progresses through the network layer by layer. Since the training time increases exponentially with the number of inputs and nodes, and polynomially with the number of training samples, refining the network to include only the most pertinent information is essential [Penny & Frost 1996]. The two main training techniques for ANNs are supervised learning and unsupervised learning. With supervised networks, the ANN is given the known output of the training set, and using that information, it compares its own calculated result, and updates the network accordingly; it learns by example. Unsupervised learning can be used for database classification, function approximation, pattern recognition, and data compression [Fausett 1994].

Backpropagation Learning Algorithm

Several different algorithms can train a network. The most commonly used training algorithm is error backpropagation, otherwise known as the generalized delta rule. Other algorithms include probabilistic, generalized regression, cascade-correlation, and conjugate descent methods. Backpropagation was originally published in a doctorate dissertation by Werbos [1974] in 1974. This work, however, did not receive the attention it deserved until it was independently discovered and published by Rumelhart *et al.* in 1986 [Rumelhart *et al.* 1986a, 1986b].

The training algorithm defines how the ANN learns, and thus depends on the purpose of the ANN [Itchhaporia *et al.* 1996]. Backpropagation is similar to the steepest descent method of optimization, where the function is the error, and the variables are the weights of the network [Fausett 1994]. The backpropagation training algorithm's name comes from the way it learns pattern recognition. First, the data is presented in the input layer, the information feeds forward through any hidden layers present, and finally yields an output response. The connection weights at the output layer are commonly initialized with small random values. The true outputs are called targets. After comparing the system's output value to the target value, the weights are tuned accordingly. The sum of squared errors cost function is commonly used, and it is described by Equation 2.3.

$$SSE = \sum_{k \in S} (target_k - prediction_k)^2 \quad (2.3)$$

In this equation, $target_k$ is the true (or desired) value of the time series at time k, and $prediction_k$ is the actual (or observed) output of the network for time k. Weight adjustments either minimize the overall error or reinforce the important weights. Then, the output error is propagated backwards through the network to the input layer to adjust the weights in each node [Livingstone *et al.* 1997]. Each cycle of presenting every training pattern to the network exactly once is called an epoch or an iteration. A step-by-step description by Baxt [1991] including the mathematical equations involved with training a feedforward backpropagation ANN can be found in Appendix C.[2]

The main drawback of the backpropagation algorithm is the possibility of finding only a local minimum, as opposed to the global minimum using the modified gradient descent methodology [Rumelhart *et al.* 1986b]. To deal with this dilemma, Penny & Frost [1996] recommend training the network several times using different initial weight settings as selected by a random number generator. The learning rate controls the size of the step taken in the direction of the gradient. Although high values speed up the learning process, if the learning rate is too high, it may oscillate around the global minimum unable to converge. Another adjustable factor used in training a network is the momentum term. The idea is to add a proportion of the previous weight-change value to the new value, thereby giving the algorithm some "momentum" to prevent it from getting caught in local minima [Penny & Frost 1996].

The most useful activation or transfer functions are continuous, differentiable, and monotonically nondecreasing [Fausett 1994]. Both the logarithmic and hyperbolic tangent transfer functions (and their derivatives) meet these requirements, thereby simplifying many ANN training algorithms. Figure 2.4 illustrates each of these functions. Due to their nonlinearity, these transfer functions have saturation values meaning that after a certain point of stimulation, the output is no longer affected [Itchhaporia *et al.* 1996]. The logarithmic, or log-sigmoid, function is commonly employed in an outcome prediction setting, and its output ranges from 0 to 1. The hyperbolic tangent (tan-sigmoid) function, however, offers a sharper transition between its output values that permits faster learning. Also, the network can learn information when a feature is absent, since the output range is -1 to +1, and the transition region crosses zero. Even with a nonlinear transfer function, a single-layered ANN can only develop linearly separable functions [Penny & Frost 1996]. This occurs because stratification into classes depends on a threshold

[2] Used with permission, see Appendix B.

that usually corresponds to the transition point of an activation function. Consequently, feeding the signals through a transfer function, linear or not, will not change the class to which the case belongs.

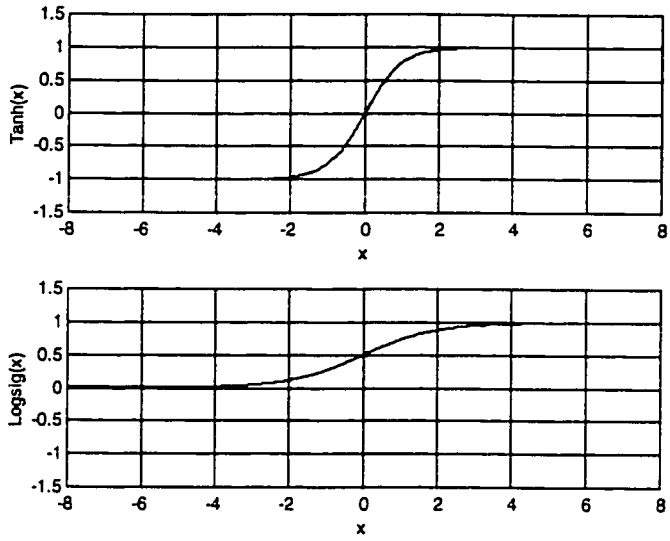


Figure 2.4: Hyperbolic tangent (Tanh) and logistic (Logsig) transfer functions

According to Richard & Lippmann [1991], the output of all ANNs can be interpreted directly as an *a posteriori* probability estimate, in other words, a probability of risk. If the *a priori* probabilities are artificially altered for training (as is intended in this thesis), then the outputs can be scaled by the ratio of the true to training prior probabilities to attain the true posterior probabilities. For example, if the prior probability of class A was artificially made 0.50 during training by sampling this class frequently and the true prior probability was 0.001, then the output of the network for that class should be multiplied by 0.001/0.50.

Weight-Elimination Cost Function

One approach to solve the problem of overfitting is to add a complexity term to the cost function. Two cost functions have proven to reduce memorization: weight-decay and weight-elimination [Weigend *et al.* 1990a, 1990b, 1991a, 1991b, Krogh & Hertz 1992]. Weight-decay was originally proposed by Hinton and Le Cun in 1987 [Weigend *et al.* 1990a]. This cost function limits the size of the connection weights, thereby penalizing large weights. The effect is a more stable network, because the output has less variance. Since weight-decay is actually contained within the weight-elimination formula, this

discussion will focus on weight-elimination. Contrary to weight-decay, weight-elimination tries to reduce the small weights to zero (in other words, eliminating the small weights from consideration). This approach is well-suited for network pruning by eliminating variables that offer little or no assistance in predicting the correct outcome [Weigend *et al.* 1990a, 1990b, 1991a, 1991b, Trigg 1997, Frize *et al.* 1997, Ennett & Frize 1998]. The small weights only add unwanted "white noise" to the model. These cost functions work best when using a large initial network structure, relatively small initial weights, and a relatively small learning rate [Weigend *et al.* 1990a, 1990b].

The penalty term in weight-elimination (the second term in Equation 2.4) "counts the number of parameters, and minimizes the sum of performance error and the number of weights by backpropagation" [Weigend *et al.* 1991a].

$$E(W) = E_0(W) + \lambda \sum_{ij} \frac{\frac{w_{ij}^2}{w_0^2}}{1 + \frac{w_{ij}^2}{w_0^2}} \tag{2.4}$$

$E(W)$ is the combined cost function that includes the initial cost function, $E_0(W)$ (typically, the SSE), and the weight-elimination term (the second term in Equation 2.4). Here, W represents the weight vector, λ is the weight-decay constant, and w_{ij} indicates the individual weight of the ANN.

The role of the weight-decay constant, λ, is to determine the relative importance of the weight-elimination term, in other words, how strongly the weights are penalized. Larger values of λ mean that a weight must be closer to zero to be considered a part of the "noise" distribution. Also, larger λ values increase the "pressure" on small weights to further reduce their size. The network is sensitive to the value of the decay constant. Choosing a value of λ that is too small will not affect the network. When λ is too large, all weights are forced to zero. A value of λ that works well for a problem that is easily identified may be too large for a more challenging situation, or the ANN may need different values of λ depending on the particular problem region it is dealing with [Weigend *et al.* 1991a]. Weigend *et al.* [1990a, 1990b] recommended starting network training with λ at zero, so that the ANN could initially take advantage of all input variables, and then slowly increase the value of λ until the network's performance begins to decline. At that point, adjust λ accordingly to optimize the performance [Weigend *et al.* 1990a, 1990b].

The scale parameter, w_0, defines the sizes of "large" and "small" weights. This scale parameter must be chosen by the user. When w_0 is small, the small weights will be forced to zero resulting in fewer large weights (i.e., weight-elimination). A large w_0 causes many small weights to remain, and limits the size of large weights (i.e., weight-decay) [Weigend *et al.* 1991a, 1991b].

Weigend *et al.* [1991a] state that the learning algorithm should "change the weights according to the gradient of the entire cost function, continuously doing justice to the tradeoff between error and complexity." Therefore, the weight update (in continuous time) will be described by Equation 2.5.

$$\frac{dw_{ij}}{dt} = -\frac{\partial E_0}{\partial w_{ij}} - 2\lambda \left[\frac{w_{ij} w_0^{\,2} \left(1 + \frac{w_{ij}^{\,2}}{w_0^{\,2}} \right) - w_{ij}^{\,3}}{w_0^{\,4} \left(1 + \frac{w_{ij}^{\,2}}{w_0^{\,2}} \right)} \right] + C \tag{2.5}$$

Although the weight-elimination technique deletes the least important input variables, this approach, among others [Chauvin 1989, Hanson & Pratt 1989, Mozer & Smolensky 1989], needs to have its "pruning" coefficients finely tuned using an iterative procedure, and the learning process is significantly slower [Le Cun *et al.* 1990]. Rather than focussing on the theory that "magnitude equals saliency," Le Cun and his colleagues [1990] suggested a theoretically justified saliency measure that they called "optimal brain damage." This theory computes the saliency measures by determining the second derivative of the cost function with respect to the input parameters. Optimal brain damage successfully eliminated three-quarters of the parameters of a practical network trained using a backpropagation ANN used to recognize handwritten digits (from 10,000 initial parameters to 2600 parameters). Although this approach was successful at handwritten zip code recognition, the experiments carried out in this thesis will apply weight-elimination, while further exploration into this technique will be left as future work.

Network Generalization

The main objective of developing models with ANNs is to apply the system to an unseen dataset, and achieve a level of performance similar to that of the training set. Generalization to a test set depends on a good balance between the complexity of the model, and the richness of the data in the training set. As the complexity of the model increases by adding more layers, more inputs variables, or more nodes, so does the possibility of overfitting the model to the training set. At present, no well-defined guidelines exist for selecting a particular ANN architecture. The number of hidden layers and hidden nodes, the

pattern of neuron connections, and the type of activation function of an optimal network are discovered using an ad-hoc approach. A set of guidelines for ANN formulation are, however, under development [Itchhaporia *et al.* 1996].

The training set should be sufficiently large to be a representative sample of the population of interest. There are two types of generalization: interpolation and extrapolation. Interpolation is the safest, since the model makes a prediction for a case that it has never seen before, but is closely related to the training data. This means that the model can base its output on other information in the database. Extrapolation occurs when the test pattern is outside the range of the training data, or within a large "hole" in the training set. Here, the model attempts to use "rules" that do not apply to the test pattern, thereby making extrapolated outputs unreliable. If the training set contains enough cases, there is less need for extrapolation [CVOR 1997]. Unfortunately, today in the medical domain, the databases rarely contain enough patient cases to cover all possible situations.

There are two commonly used criteria for stopping ANN training on a training set: setting bounds on the change in error and early stopping. The idea behind the first approach is to stop training once the change in error is less than a predetermined value. Although the advantage of this technique is that the stopping criterion does not depend on the test set, by the time the network reaches the point where the change in error is small, the network may have overfitted to the training set, thereby reducing the model's ability to generalize.

Early stopping is a commonly used approach to finding the best-performing ANN model. It requires at least two subsets of a database: one for training and the other for testing. Early stopping means that training is stopped once the test set error rate begins to rise. This approach seems to prevent both overtraining and overfitting [Livingstone *et al.* 1997]. Below is a "recipe for success" for developing a useful ANN model using early stopping as described by the Consortium for Virtual Operations Research (CVOR) on a West Virginia University webpage [CVOR 1997].

"1. Divide the available data into training and test sets.
2. Use a large number of hidden units.
3. Use very small random initial values.
4. Use a slow learning rate.
5. Compute the test set error rate periodically during training.
6. Stop training when the test set error rate starts to go up."

To look at the network performance on another set to choose when to stop training is sometimes considered "cheating," so it is generally recommended to have a third set (from the same sample) to

validate the network. After determining the best performance of the training set on the test set, its generalization performance is evaluated using the validation set. The main advantage of early stopping is that it is fast and easy to apply to a network where the number of weights far exceeds the sample size.

Medical researchers have used ANNs to predict surgical and ICU outcomes, to predict lengths of stay in the ICU following coronary artery surgery, to recognize the presence of an acute myocardial infarction, coronary artery surgery outcomes, and several other medical outcomes with varying degrees of success [Baxt 1991, Frize *et al.* 1993, 1995, 1997, Tu & Guerriere 1993, Buskard *et al.* 1994, Orr 1997, Trigg 1997, Frize *et al.* 1997]. It is generally accepted knowledge that ANNs are particularly useful when there are nonlinearities in the data [Penny & Frost 1996]. Since medical data are believed to often contain nonlinear relationships between the input variables and the outcome, ANNs are a logical choice for analysis of the medical domain [Baxt 1994b].

Some researchers have reservations about using ANNs. The information the ANN learns is contained within the interlayer connection weights, so their complex mathematical structures are often reduced to the idea of a "black box." Since the knowledge is distributed across the weights, determining it exactly what was learned by the ANN is difficult [Forsström & Dalton 1995]. The computational complexity of these algorithms is enormous and makes ANNs appear overwhelming. However, every decision made by the ANN is based on mathematical reasoning and computer algorithms created to simulate the learning and decision processes that occur in the human brain. The following section outlines some common types of ANNs used in medical research.

2.5 (c) i. Multilayer Perceptrons

Lippmann & Shahian [1997] applied the patient cases from the STS database to single-, two- and three-layer multilayer perceptron (MLP) networks, otherwise known as feedforward backpropagation ANNs. These MLP networks were trained using stochastic gradient descent with early stopping. The STS database contained 80,606 patients who underwent CABG surgery in 1993. The training and test sets were randomly split approximately in half. The outputs were separated into six mortality risk categories using 36 predictor variables: 0-2.5, 2.5-5, 5-10, 10-20, 20-30, 30-100 percent. The mortality rate for the training set was 3.4 percent [Lippmann & Shahian 1997].

The networks showed poorer calibration for the two- and three-layer MLPs, however, all classifiers had approximately the same performance with all ROC curves being about 76 percent. Grover

et al. [1995] expressed an opinion that an ROC value of greater than 80-85 percent for CABG mortality may never be achieved. Warner [1997] suggested that these poor results obtained by Lippmann & Shahian [1997] may be attributed to the absence of complex nonlinear relationships among the variables presented to the networks.

2.5 (c) ii. Probabilistic Neural Networks

Orr [1997] attempted to predict mortality following open-heart surgery using a probabilistic ANN. In 1990, Specht [1990] developed the probabilistic neural network (PNN) which uses the Bayesian pattern classification technique. Although Bayesian theory is familiar to medical researchers, few have employed PNNs in medical situations. Where PNNs have been used, researchers have achieved good performance (cardiac surgery ICU length of stay by Orr *et al.* [1995a], and prediction of chronicity in a surgical ICU by Buchman *et al.* [1994]). Unlike backpropagation ANNs that require iterative learning, PNNs only need a single pass of the training set to learn the required information and are less vulnerable to memorization. Orr [1997] states that the training time is drastically reduced with PNNs, however, this statement is misleading. Although a PNN only requires a single pass to train the network, it is time-consuming to estimate the bandwidth of the probability density function [CVOR 1997]. A smoothing factor defines the width of the calculated Gaussian curve for each probability density function. Small changes in the smoothing parameter do not have major effects of the PNN's accuracy, however, inappropriate calculation of this factor may prevent the network from generalizing. The objective of the smoothing factor is to reduce the effect of variables that have only a minimal correlation with the desired outputs [Orr 1997].

Orr used a commercially available PNN (NeuroShell 2 – Ward Systems Group, Frederick, MD) to perform his experiments, and preprocessed the data to reduce the influence of outliers. To avoid overfitting, Orr minimized the number of variables used in the final model. The seven chosen variables were: age, gender, ejection fraction, IABP placed preoperatively, reoperation, not CABG, and creatinine level. Orr selected these variables from a list of those frequently used by other researchers only if the variables were easily collected, quantifiable, and unambiguous (i.e., no subjectivity) [Orr *et al.* 1995b]. The database included 1477 patients who underwent cardiac surgery between 1991 and 1994. Postoperative death was defined as death during hospitalization. The mortality rates for the training, test and validation sets were 4.65, 4.21, and 4.11 percent, respectively [Orr 1997]. The PNN achieved an overall accuracy of 91.5 percent in the training set, 92.3 percent in the test set, and 88.2 percent in the

validation set. The areas under the ROC curves were 0.72, 0.81, and 0.74, for the training, test, and validation sets, respectively [Orr 1997].

The fact that PNNs require storage of the entire training set in memory can potentially account for its lack of popularity among researchers, however, the increasing storage capabilities of today's computers should eliminate this problem. Although PNNs have been identified as less suitable for the higher-dimensional classification problems that are commonly encountered in medical decision making [Orr 1997], they can convey information about how similar a test pattern is to the training data (i.e., has a high density). A low density might indicate that the network is extrapolating information to determine the outcome, meaning that the output is less reliable [CVOR 1997].

Chapter 3: Problem Formulation and Methodology

This chapter establishes a foundation that guides the reasoning behind the chosen path for solving the research problems. The early sections of the chapter identify concerns that arise with coronary artery surgery risk models, and propose solutions for these issues. The next section highlights important results regarding certain aspects of ANN behaviour. A detailed description of the SFHI database, the selection of the initial input variables, and data manipulation follows this discussion. Finally, the chapter rounds out with a comparison of two methods of network pruning to reduce its complexity, since a model that generalizes well with the fewest input variables is preferred.

3.1 Challenging Issues of Coronary Artery Surgery Databases

The greatest debate in the search for a universal CABG surgery risk model is: what are the risk factors? After reviewing the existing risk models, one discovers that no two models find the same set of statistically significant risk factors for open-heart surgery [Edwards *et al.* 1994a]. Several factors, however, are either statistically found or clinically known to be preoperative risk factors, and, thus, are common to many risk stratification models. Such factors include the degree of cardiac dysfunction, urgency of operation, advanced age, gender, incidence (i.e., first operation or first reoperation), and the presence of comorbidities [Higgins *et al.* 1992, Shroyer *et al.* 1999]. So far, no one has developed a system that is universally applicable [Turner *et al.* 1995]. The following sections pinpoint the difficulties that arise when dealing with a CABG surgery database.

Understanding the important analytical aspects of the coronary artery surgery risk models in the literature is a complicated matter, given the variability of the characteristics of each model [Turner *et*

al. 1995]. Inter-institutional differences regarding coronary artery surgery databases that hinder comparability include:

- the mortality rate;
- the length of the period (and years) of data collection;
- the patient set to which the model caters (i.e., just CABG patients, CABG plus valve procedures, all open-heart surgeries, etc.); and
- the model development process.

All these variations in databases account for the difficulties encountered when trying to develop a model that performs well for various institutions.

3.1 (a) Mortality Rate

Different definitions of "death" will cause variations in mortality rates recorded at individual institutions. "Death" has been interpreted as operative death, in-hospital death, and death within 30 days of surgery, among others. A broader definition of death likely leads to a higher mortality rate. Applying a risk model from literature to a database with a different definition of "death" may result in poorer performance. A risk model will predict "death" as it was originally defined, and not as that particular hospital defines it.

The primary obstacle faced by medical researchers attempting to develop a coronary artery surgery mortality risk model is the low prevalence of death [Ennett *et al.* 1999, full article included in Appendix D[3]]. Although survival is desirable from a medical point of view, the low mortality rate makes it difficult for researchers to identify the risk factors. Usually, the mortality rate for heart surgery is less than 5 percent, however, higher mortality rates have been reported at certain institutions, including the Beth Newark Israel Medical Center [Parsonnet *et al.* 1989]. As expected, the very highest-risk patients are usually screened out of consideration for surgery by the surgeon. Most CABG surgery risk models can easily identify good outcomes in low-risk patients because of the abundance of data representing this situation. Despite this, a lack of sufficient data on patients suffering an adverse event prevents these risk models from being useful as clinical aids. A patient should not be withheld treatment even if the mortality risk model suggests a high-risk outcome because the models are not sufficiently accurate. Nevertheless, a high-risk score can highlight a patient to ensure additional attention during the recovery phase. Theoretically, the higher the incidence rate, the easier it is to develop a well-fitting model. The mortality rate of the training set can also affect the performance of the risk model. Baxt & Skora [1996]

[3] Used with permission, see Appendix B.

found that training a model with a higher-than-normal prevalence of a particular outcome improved the model's performance on the test group. Their findings are discussed in more detail in Section 3.2.

3.1 (b) Period of Data Collection

The period over which the data were collected is particularly relevant. If the data span too many years, the changing patient population and hospital procedures may affect the results [Turner *et al.* 1995]. If there were modifications to the surgical and/or pre- and postoperative procedures performed over the period of data collection, there may also be a marked change in the mortality rate. Clark *et al.* [1994] proved that the surgical profile of CABG patients has noticeably changed over a ten-year period based on the STS' National Cardiac Database for CABG from 1984-93. Most importantly, they observed a dramatic decrease of 17.5 percent in the lowest risk patients, and an increase in the higher-risk groups (for patients in the risk categories of 5-10, 10-20, 20-30, and 30-50 percent, their risk of mortality increased over the decade by 6.2, 9.1, 1.4, and 1.1 percent, respectively). Although the risk profiles of typical CABG patients changed over that particular decade, the actual observed operative mortality did not show an appreciable increase. The steady mortality rate, despite the increase in number of higher-risk patients, may be attributed to technologic, pharmacologic and operative advances [Clark *et al.* 1994].

Another problem arises when researchers compare models that were designed using data from different time periods. As discussed above, patient profiles, surgical procedures, pre- and postoperative care, and the list of relevant risk factors are subject to change over time. Edwards *et al.* [1988] recognized this issue, and noted that due to the changing environment, risk models developed even a few years before may not be applicable to the current data. Also, when Orr *et al.* [1995b] compared four severity-adjusted models for predicting CABG surgery mortality (Cleveland Clinic, New York State model by Hannan *et al.* [1994], Parsonnet, Northern New England Cardiovascular Disease Study by O'Connor *et al.* [1992]), they noted that the Parsonnet model consistently predicted higher mortality rates than those actually observed. To account for this discrepancy, they suggested it might not be representative of current practice since the Parsonnet model was created much earlier than the others. These authors [Orr *et al.* 1995b] note also that the Parsonnet database had a significantly higher mortality rate than most institutions (8.9 percent). Again, comparing models that do not share similar patient and hospital profiles does not accurately demonstrate their performance.

3.1 (c) Model Development Process

The lack of a uniform approach to model development also makes the process of evaluating the risk stratification models against one another difficult. As of yet, no guidelines have been developed to improve the model selection approach for coronary artery surgery models. The differences in the model building process include: the patient dataset for which the model is designed; the number of patient cases upon which the original model was based; risk factor definitions; how the variables were selected; how the risk factor weights were chosen; the number of identified risk factors; how the model was evaluated; and whether the risk model has been tested at institutions other than the originating institution.

Applying risk stratification models to the patient dataset for which it was designed (i.e., CABG only, all open-heart surgeries, etc.) is important, otherwise the accuracy of the outcome predictions may be compromised. The patient population (i.e., the number of higher-risk surgeries performed, disease prevalence, etc.) plays an important role in establishing the mortality rate at an institution. Some types of cardiac surgery are riskier than others. Thus, the type of surgical procedure needed by patients influences the mortality rate.

Small sample sizes do not reflect a random sample from the general population. This usually means that there are not enough cases to adequately represent all of the possible combinations of risk factors. On the other hand, large populations may cause some variables to appear to have an overinflated impact on the outcome of interest. A certain factor may seem significant because the population is so large, but may not necessarily be a risk factor. Also, large databases tend to comprise data from several institutions. The larger institutions can introduce a bias, because obviously they would be submitting more data than smaller institutions. Nevertheless, as previously stated, Edwards *et al.* [1994b] suggest that this bias would likely cancel out over time as more small practices are entered into the database population.

The approach that the researchers take for the selection of risk factors is sometimes arbitrary. In some cases, the variable selection was based solely on expert opinion (i.e., Edwards *et al.* 1988). Usually, the variables are selected using univariate and multivariate analysis, as well as clinical intuition. Unfortunately, authors rarely identify which risk factors were chosen by which method. Risk factor weights for additive models are sometimes assigned subjectively as well, or the coefficients may simply be the odds ratio. As the number of identified risk factors increases, so does the chance of overfitting the model to the training data. The number of factors in coronary artery surgery risk stratification models

ranges from seven to more than 30. A greater number of risk factors compromises the generalizability of the model and increases the cost of collecting patient data.

A lack of universally-accepted risk factor definitions compromises the comparison of different risk models. Similar to the problem with the definition of death, variations in the definitions of certain risk factors also affect their prevalence.

The poor generalizability of most CABG risk models may be attributed to the absence of some important risk factors whose influence has not yet been recognized. Another possibility is that relevant risk factors may have been selected, but their impact on the model is not obvious because of their infrequent occurrences in the database [Lippmann & Shahian 1997]. Perhaps as institutions continue to collect additional information on their cardiac patients, new risk factors will appear. Inter-institutional differences such as differences in medical procedures or hospital policies may also affect the parameters in the mortality risk models.

Researchers should be diligent and use statistical analysis of their model's performance with care. Classification accuracy can be biassed by disease prevalence and by manipulation of the cutoff value, so overall accuracy may be used to measure the system's performance (defined as true positives plus true negatives divided by total sample size). It often seems that the influence of the prevalence of the desired outcome is not adequately explored. It is generally accepted that ROC curves offer the least biassed point of view with respect to the mortality *a priori* probabilities, although prevalence is still a factor [Buchman *et al.* 1994, Forsström & Dalton 1995, Orr 1997]. The development of a completely unbiased performance measure would possibly enhance model evaluation for highly skewed data, such as coronary artery surgery mortality data.

Although developing a coronary artery surgery risk model that generalizes well to other institutions is difficult, testing the model performance on data from other hospitals and practices is an important technique for validating the model. All risk models must be validated on a test set that was not used to develop the model, but comes from the originating institution. A validation set from another institution is useful, but few researchers perform this test on their own model. Validation set testing usually occurs when another research group develops a model, and compares their own model with those in the literature.

There is a potential negative impact of using risk stratification methods to assess an institution's performance. The health care practitioners may be tempted to modify their reporting practices to accommodate a higher mortality rate at the institution. These "report cards" may cause an increase in rejected procedures, or an institution may hesitate to accept higher-risk patients for fear of having an overall higher than normal mortality rate. This would result in an "outmigration" of higher-risk patients [Lippmann & Shahian 1997]. These issues should be thoroughly investigated before the implementation of a definitive CABG risk model at any institution.

3.2 Possible Solutions

Several authors have suggested solutions to the challenges identified in Section 3.1. In this section, potential remedies for each problem outlined in the previous section are introduced.

3.2 (a) Modifying the Database

It is well known that the performance of the model depends strongly on a balance between the network complexity and the information contained in the training samples. Researchers have referred to a minimum-acceptable number of cases per weight in a model. A minimum of 6 to 10 samples per input variable is generally acceptable [Weigend *et al.* 1990a, 1990b, Forsström & Dalton 1997]. Obviously, a larger database provides more examples of the underrepresented outcome despite its low prevalence. Thus, waiting until a sufficient number of cases have been collected is one possible way to obtain a database with enough representative samples. For example, recall that although the mortality rate of the 1996 STS database was only 3.2 percent, there were nearly 2800 cases of nonsurvivors of CABG surgery [Shroyer *et al.* 1999]. For individual institutions, however, collecting such amounts of data is impossible. Therefore, other alternatives should be explored.

Another approach would be to include information about patients who are denied surgery. An unsuitable surgical patient likely has very advanced co-morbid conditions that would make anesthesia dangerous, poor flow in the distal portion of the native vessel that cannot be rectified with endarterectomy, and/or very poor ejection fraction. It is likely that patients who do not survive the heart surgery would have similar characteristics to those deemed unsuitable for the surgery in the first place. This information could be combined with that of the nonsurvivors, thus "theoretically" increasing the number of cases of potential nonsurvivors. By increasing the number of cases representing the patients

[illegible] hypothesized that a more accurate risk model could be [illegible] few hospitals collect this data, so until data from a [illegible] are examined testing this is not possible.

[illegible] data to represent [illegible] scenarios is not feasible, and there [illegible] were denied surgery, then a possible solution is to artificially alter the [illegible] database. This feat may be achieved by either reducing the number of [illegible] increasing the number of "nonsurvivors." Some researchers have alluded [illegible] Smith 1993, Ohno-Machado *et al.* 1998], but to the [illegible] seems to have tested the concept on real medical data.

In an attempt to improve machine learning performance when using a dataset of patients presenting themselves to the emergency room suspected of having myocardial infarction (MI), Ohno-Machado *et al.* [1998] randomly removed patient cases, and observed the effect on the areas under the [illegible] Ohno-Machado *et al.* [1998] suggested that this technique removes "redundant" cases in order to speed up the learning time for the ANN. Starting with 700 patients, they concluded that as many as 90 percent of the cases could be randomly removed without affecting the area under the ROC curves. In this situation, all cases (those patients with MI and those without) were vulnerable to being randomly removed. Since the medical problem here was MI, the distribution of the cases is similar to Baxt & Skora [1996] with about 7 percent of the cases representing those patients who did indeed have an MI (typical incidence rate of inner city hospitals [Baxt & Skora 1996]). The successful performance of the ANN on this database, despite the low prevalence of MI, suggests that those patients had certain characteristics that easily identified them as having an MI.

Research by Ennett and Frize [1998] showed that the performance of an ANN using the DECH ICU database was inconsistent when the percentage of cases of the nondominant output was near 15 percent or less (i.e., sometimes the ANN would work and other times it would classify like a constant predictor). This implies that a representation of at least 20 percent would include a "factor of safety." Based on this theory of artificially altering the data distribution, duplicating the nonsurvivor cases of the SFHI coronary artery surgery database until they represent at least 20 percent of the database, or randomly deleting the survivor cases until the representation of the nonsurvivors is at least 20 percent are possible approaches. Of course, this technique would increase the sensitivity of the model at the cost of specificity.

Baxt & Skora [1996] trained an ANN on a set with an MI prevalence of 34 percent (120 patients with MI out of 350) while the test set had a frequency of MI of only 7 percent. Baxt and Skora did not exactly describe how the selection of training cases was made, however, no mention of preferential case selection was mentioned. The network performed significantly better when trained on a population of patients with a higher prevalence of MI than a network trained on a group with a low prevalence (and thus closer to reality) of MI [Baxt & Skora 1996]. It is interesting to note that they did not use ROC curves to measure the performance of the networks. Ninety-seven percent of the time, the raw output values for this network were either less than 0.01 or greater than 0.90, so an ROC curve would not offer much information about the model's operating point [Baxt & Skora 1996].

3.2 (b) Limiting the Time Span of Data Collection

Because certain characteristics can dramatically affect a patient's risk of dying, being aware of changes in the patient population over the years of data collection is crucial. This explains why developing a model based on just a few years of data as opposed to a database containing patient profiles over a decade may be preferable. A more homogenous dataset will result, and this approach will also reduce the impact of changes to the surgical, pre- and postoperative procedures. The less variation there is between the patient cases, the easier it is to develop a well-fitting model. Despite the recommendation to use data spanning fewer years, an in-depth investigation into the statistical profile of any database is essential before the model development process begins. The STS committee is implementing this idea, and has begun developing models for each separate calendar year to accommodate the temporal changes in the patient population [Edwards *et al.* 1997].

3.2 (c) Developing Guidelines

The easiest way to reduce the variability between CABG surgery databases is to develop a set of data collection techniques that are universally-acceptable. The STS committee has designed such a guideline for the institutions that participate in their studies. Being the largest national-based database of this type in the world, following the lead of the STS committee seems appropriate for other organizations and other countries. This set of instructions should be universally available to guarantee that inter-institutional evaluations are not comparing "apples and oranges." Allowing all institutions to use their guideline would most likely benefit the STS committee. American institutions that are not currently involved could test the STS models on their own data that was collected according to the instructions outlined by the STS committee. If the model performed well on their data, the institution or practice might be more inclined to join such a national data collection effort. International organizations

may be stimulated to start collecting their own national database, following the STS guidelines to ensure uniformity. This approach would more clearly delineate any real differences between coronary artery surgery patients from different countries, an interesting comparison to make, and perhaps validate the work done by the STS committee.

Typically, variables used in risk stratification models for coronary artery surgery are categorical, and refer to the presence or absence of certain comorbidities or medical history. Warner [1997] suggested using variables other than just risk factors, however, medication and mechanical support can normalize physiological characteristics, thus making interpretation difficult [Turner *et al.* 1995]. Currently, the SFHI database does not contain physiological information, but the SFHI is in the process of collecting a database of pre- and postoperative patient characteristics that could provide useful information about operative mortality in the future [Shaw 1999].

A second point to consider about medical databases is the way the risk model results are presented. Former MIRG graduate student, Trigg (nee McGowan) developed a reporting guideline for medical researchers [McGowan *et al.* 1996]. Although her conference paper was geared towards ANN research, the ideas presented in the paper are generalizable to all types of model development. Trigg suggests that certain details about the dataset under investigation be reported, such as the sample sizes, the number of variables and types, and the classification rate of a constant predictor. The specific characteristics of the modelling approach should be outlined. Trigg referred to details about the ANN, however, statistical techniques should also be described with precision such as the level of significance to retain or reject input variables in the model. Finally, the results should be carefully reported by including the classification rates on the training and test sets, the sensitivity and specificity of the model, ROC curves describing the C-index, and a comparison with other benchmark models [McGowan *et al.* 1996]. These guidelines for reporting experimental results include the most important information that can be used to compare one model with another in the literature. Thus, to make the results from this thesis interpretable by other medical researchers, Trigg's reporting guidelines were followed as closely as possible in Chapter 5 of this thesis where the experimental results are presented.

3.3 Significant Advances in ANN Research

Although most medical applications of ANNs focus solely on the network performance rather than the particular attributes of ANNs or attempts to improve the performance, there has been a small amount of work in this area. The following sections discuss advances made on studying the behaviour of ANNs in the medical domain.

3.3 (a) Weight-Elimination Variables

Before analysing the current CABG patient database, an investigation of the ICU database from the DECH revealed some interesting results. As a continuation to the work of Trigg [1997] and Buskard [1994], a series of experiments using a feedforward backpropagation ANN were performed to investigate the relationship between the number of input variables of the postoperative database and a dichotomous output variable (less than 8 hours of mechanical ventilation, or greater than or equal to 8 hours of mechanical ventilation) [Frize *et al.* 1998].

Using high-low node representation and weight-elimination on a single-layered network from the postoperative database, the six variables with the largest weight values after sufficient training were presented to the ANN. The results indicated that by reducing the number of variables from 51 to the six most important variables (as selected by weight-elimination), approximately the same level of performance was achieved in fewer epochs, and therefore in a shorter period of time [Frize *et al.* 1998].

The fact that reducing the number of variables did not dramatically affect the network's overall performance is important. Requiring fewer input variables for estimating outcomes will reduce the time required for data collection, and hence the cost of data collection, as well as the time needed to obtain results for a new patient. The most significant discovery here, however, is that it appears that the ANN can select which variables have the most influence on the outcome variable. Weigend *et al.* [1990a] also interpreted the performance of the weight-elimination cost function as having the ability to eliminate the least important variables. Hence, an ANN with weight-elimination may be used as an alternative variable reduction technique that selects the most important variables without researcher bias.

3.3 (b) Statistical Deduction of Minimum Number of Cases Required

In another set of experiments using the DECH ICU database and ANN architecture outlined in Section 3.3(a), the objective was to determine the minimum number of sample patterns needed by the ANN to achieve a satisfactory level of performance[4] [Ennett & Frize 1998]. This time, the same 20 input variables were employed in each experiment, while several different dichotomous output variables were used. The output variables had different distributions of the dominant outcome ranging from 52.0 to 98.3 percent.

The performances of ANN trained on both the postoperative and nonpostoperative databases were compared. Using a linear regression of the ANN results, an approximate percentage of sample cases that the ANN needed to perform better than a constant predictor was found. For the postoperative patients, the overrepresented class could not be more than 85.9 percent of the training cases. When the dominant output class represented more than 83.5 percent of the nonpostoperative sample cases, the constant predictor outperformed the ANN model. Hence, in order for the ANN to perform satisfactorily, approximately 15 percent of this ICU database must contain the underrepresented-class sample cases. These experiments proved that the ANN requires a certain amount of information about the underrepresented class to develop a good model.

3.3 (c) Clinical Importance of Variables

W.G. Baxt was a forerunner in the application of ANNs to the medical domain. Baxt has published several papers outlining his progress using feedforward backpropagation ANNs to detect the occurrence of acute MI in an emergency care setting. Although the clinical setting is different, the discoveries that Baxt made with respect to the power and functions of ANNs are particularly significant for medical domains. In his 1991 paper [Baxt 1991], Baxt achieved impressive results when the ANN had a significantly higher sensitivity and specificity compared with the emergency department physicians (sensitivity 97.2% vs. 77.7%, and specificity 96.2% vs. 84.7%). Baxt suggested that a possible reason for the improvement was that the ANN can possibly identify certain relationships between the input data that are nonlinear, and these may not be apparent when using other approaches, perhaps not even to experts.

Later, Baxt identified an important finding: some input variables that were clinically thought to be of little importance for identifying acute MI showed significant positive effects (favouring diagnosis)

[4] Here, "satisfactory performance" is defined as "an improvement over a constant predictor."

when using an ANN [Baxt 1994a]. This discovery is particularly interesting because it gives strength to Baxt's hypothesis that the ANN could recognize relationships that are not evident using other paradigms. According to Baxt, in most MI cases, the relationship between clinical data and the presence of MI is linear. In the case of a linear relationship, the ANN cannot improve the diagnostic accuracy much, since it will also relate the variables linearly. The power of an ANN is seen when investigating patients who are in the minority, those whose clinical symptoms do not appear to be linked directly to the disease. It is here that the ANN may afford a greater diagnostic accuracy by recognizing these unapparent relationships. The results found here may also indicate that a potential loss of information occurs when analysing data with strictly linear techniques. As discussed in Chapter 2, an ANN can adapt itself to the database environment and use the most appropriate modelling approach. Baxt's findings emphasize the modelling power of ANNs to recognize relationships that would otherwise go unnoticed. ANNs may aid in the discovery of new variables whose influence had not yet been considered.

3.4 The SFHI Cardiac Patient Database

To understand the situation under investigation better, a general demographic comparison with other institutions followed by a detailed description of the SFHI database are necessary. Table 3.1 presents the prediction model characteristics discussed in Sections 3.1 and 3.2, such as the number of years that the data collection spans, the size of the training set, the number of risk factors used in the model, how many institutions participated in each study, how the researchers defined "mortality" in each case, and the crude mortality rate for each study. These characteristics are compared for each of the models described in Chapter 2 (additive, statistical and ANNs). Knowing this information is useful when applying the risk models from the literature to a new test set (i.e., when researchers compare the performance of models from the literature on their data against a model developed based on their data – a "generalizable" model versus an institution-specific model). Recall that a comparison is valid when the population upon which the model was developed and the test set are similar. Note the different characteristics of the eight models in Table 3.1. No two models share the same profile.

Table 3.1: Profiles of prediction models in the literature

Model	Years of data collection	Number of patients (training set)	Number of variables	Number of institutions participating	Mortality definition	Crude mortality rate
Parsonnet Model [Parsonnet *et al.* 1989]	1982-1987	3500	17	1	within 30 days of CABG	8.9
Cleveland Clinic [Higgins *et al.* 1992]	1986-1988	5050	13	1	in-hospital or within 30 days	2.5
STS1 [Edwards *et al.* 1988]	1984-1987	300	21	1	within 30 days of CABG	4.75
STS2 [Edwards *et al.* 1994a]	1984-1990	39464	20	> 100	operative	3.2
SFHI Bayes' [Pliam *et al.* 1997]	1985-1994	2842	29	12	in-hospital	4.0
SFHI Logistic Reg [Pliam *et al.* 1997]	1985-1994	2842	13	12	in-hospital	4.0
MLP network [Lippmann & Shahian 1997]	1993	40303	36	> 100	operative	3.4
PNN [Orr 1997]	1991-1994	1477	7	1	in-hospital	4.65

Table 3.2 compares the population demographics of several commonly-used risk factors for CABG mortality (mean age, gender, percentage of blood ejected from the heart, whether the surgery was a reoperation, whether the patient underwent a preoperative IABP, presence of diabetes or COPD, and the status of the operation) of the SFHI with other institutional data that were available in the literature.

Table 3.2: Population demographics of various institutions

Risk factor	SFHI	Cleveland Clinic [Higgins *et al.* 1992]	STS2 [Edwards *et al.* 1994]	New England[5] [O'Connor *et al.* 1992]
Mean age, years	63.8	NA	62.5	63
Female	24.7	20.6	24.6	26.8
EF < 30 percent	9	NA	9.5	NA
EF <40 percent	NA	NA	NA	9
Reoperation	10	18.5	6.7	6.1
Preoperative IABP	4.2	NA	5.9	NA
Diabetes	22.8	17.2	17.9	18.1
COPD	13.2	7.5	3.2	11.2
Emergency operation	5	3.1	NA	6.6

* EF=ejection fraction of blood from heart, NA=not available

Table 3.2 shows that most characteristics of the SFHI's patient population are similar to the other American hospitals presented here. According to the data, the SFHI has higher rates of diabetes and COPD than the institutions involved with the Cleveland Clinic, the Society of Thoracic Surgeons and the New England study group. A model should be tested on a dataset from either the same population or a population with similar demographics. If the training and test groups have dramatically different

[5] New England refers to the Northern New England Cardiovascular Disease Study Group [O'Connor *et al.* 1992].

population characteristics, then the model may not be valid on the test set. Tables 3.1 and 3.2 reemphasize the care that should be taken when interpreting the performance of a risk model from a different institution.

The SFHI's cardiac database has 7050 patients who underwent all types of open-heart surgery between January 1, 1985 and June 30, 1994. All of the variables available for this research are categorical except the patient's age and the date of surgery. The only surgical cases included in this analysis were those patients who underwent CABG surgery, CABG plus valve surgery, or CABG plus repair surgery. This is the same set of cases used by Pliam *et al.* [1997] in their research. Figure 3.1 shows the breakdown of the database by surgical type. The total number of cases in this reduced dataset was 6325. Out of those 6325 cases, there were 248 deaths giving an overall mortality rate of 3.9 percent.

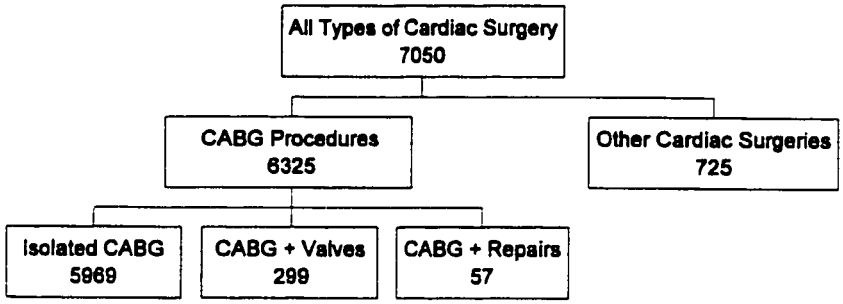


Figure 3.1: Breakdown of SFHI database by surgical type

Turner *et al.* [1995] commented that most studies span many years. These studies, typically, do not take into account the changing patient profiles and mortality rates over time, or changing surgical procedures and patient management techniques. This observation suggests focussing on a smaller time profile. Dr. R.E. Shaw from the SFHI mentioned that some technical advances were introduced at the hospital during 1990-91 that were likely felt by patients in 1992 and afterwards. These technical advances to improve surgical outcomes included the implementation of better cardiopulmonary bypass machines used during surgery, increased efforts to extubate patients as quickly as possible following surgery, increased prophylactic use of antibiotics to prevent infection, and a more aggressive approach to improving mobilization of patients after surgery [Shaw 1999]. To maintain the most homogeneous database possible, the surgical and hospital procedures should not change over the duration of data collection. This information prompted a review of the annual mortality statistics over the period of data

collection. Figures 3.2 and 3.3 reveal the annual sample sizes and mortality rates for the SFHI database for all years of data collection.

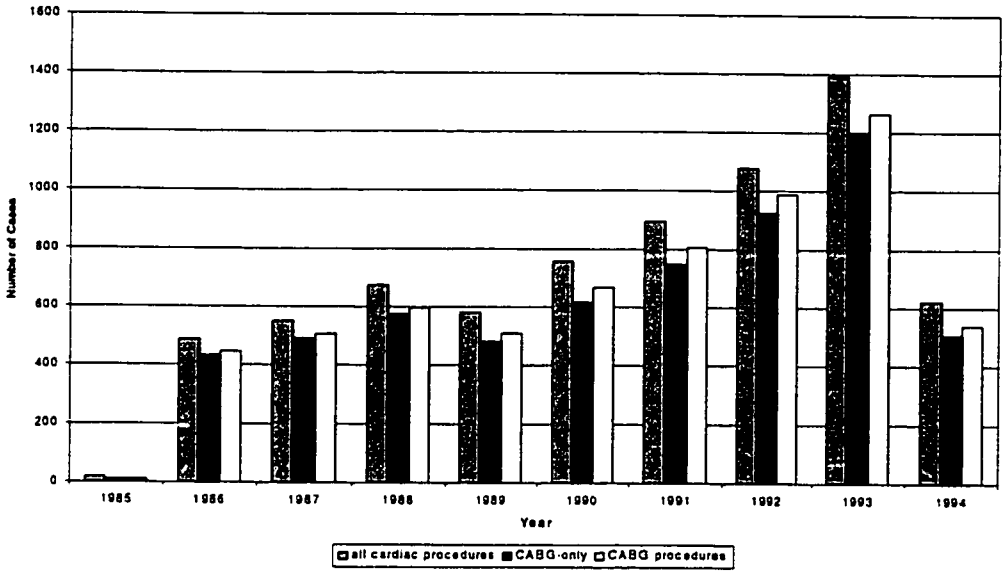


Figure 3.2: SFHI cardiac database annual profile (1985-94)

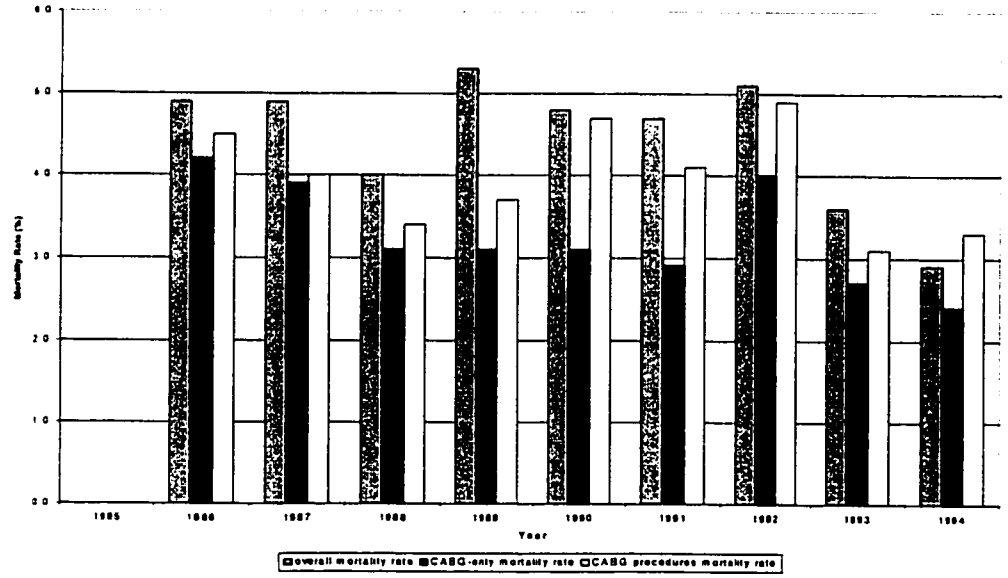


Figure 3.3: SFHI cardiac database annual mortality rates (1985-94)

These figures show the annual mortality rates, the distribution of patients who underwent CABG surgery, and those who had combined operations of CABG plus valve or repair surgery. The mortality rate for patients with just CABG surgery is less than that of the overall group of heart surgery patients. The analysis showed that only nine patient cases were collected in 1985, so these cases were deleted because they were not representative of the CABG surgeries in 1985. A minor increase in the annual mortality rate occurs in 1992, which is followed by a significant drop in 1993 and 1994. The higher mortality rate in 1992 might be attributed to less stringent patient selection by the surgeons. Since the surgeons at the SFHI regularly receive feedback on their performance, they recognize that the patients requiring surgery are sicker. Therefore, the surgeons may have to accept more difficult cases, thereby operating on patients who have a higher risk of mortality, and ultimately affecting the overall mortality rate of the institution. Once the pattern of increased mortality due to the leniency of admission is recognized, the surgeon must begin choosing patients based on more stringent acceptance criteria. The fluctuating mortality rate due to patient selection criteria is common at many institutions, and therefore is not unique to the SFHI [Shaw 1999]. Shaw's comments confirm that the patient cases from 1992-94 are from a different pool of data, and thus these cases were dropped from consideration. Thus, a more homogeneous database spanning the years of 1986-1991 with 3526 patient cases remains. Within this dataset of CABG, CABG plus valves, and CABG plus repair surgery patient cases, there were 143 deaths giving an overall mortality rate of 4.1 percent.

Table 3.3 contains the variable list that was initially considered for experimentation. The variable list comprises those factors chosen by Pliam *et al.* [1997] by univariate analysis, as well as other variables that are commonly used in other models. The initial variable list contains 39 input variables and three possible output categories (DEATH, ICUSTAY, HSPSTAY).

Table 3.3: Initial variable list for thesis experiments

SFHI Variable	Definition	Values	Type*
SURGDT	date of surgery	ranges from Jan 1/86 to Dec 31/91	C
AGECAL	patient's age	ranges from 20-92 years	C
FEM	female gender	1=female, -1=male	B
ETHNIC	ethnicity	1=Caucasian, 2=Black, 3=Asian, 4=Native American, 5=Hispanic, 6=Filipino, 7=other	N
OPMVD	mitral valve disease operation	1=yes, -1=no	B
OPAVD	aortic valve disease operation	1=yes, -1=no	B
EMERURG	emergent/urgent priority for surgery	1=yes, -1=no	B
PTCA	failed PTCA prior to surgery	1=yes, -1=no	B
REOP	reoperation	1=yes, -1=no	B
RENLDX	renal disease	1=yes, -1=no	B
DISVAN	ventricular aneurysm	1=yes, -1=no	B
DISLMAIN	left main disease	1=yes, -1=no	B
EJFRAC	ejection fraction	1=normal, 2=moderate, 3=severe	O
DISMR	mitral valve regurgitation	1=yes, -1=no	B
DISAS	aortic valve stenosis	1=yes, -1=no	B
DISTRIV	tricuspid valve disease	1=yes, -1=no	B
HYPERTEN	hypertension	1=yes, -1=no	B
PREIABP	pre-operative intraaortic balloon pump (IABP)	1=yes, -1=no	B
MI	previous myocardial infarction (MI)	1=yes, -1=no	B
EVOLMI	evolving MI	1=yes, -1=no	B
MIONSET	onset of MI	1=remote, 2=recent	O
HXCHF	history of congestive heart failure	1=yes, -1=no	B
CHF2	current congestive heart failure	1=yes, -1=no	B
UNSTANG	unstable angina	1=yes, -1=no	B
CVDDX	cerebrovascular disease	1=yes, -1=no	B
PVD	peripheral vascular disease	1=yes, -1=no	B
TVD	triple vessel disease	1=yes, -1=no	B
OBESE	obesity	1=yes, -1=no	B
SMALL	small stature	1=yes, -1=no	B
COPD	chronic obstructive pulmonary disease	1=yes, -1=no	B
DIAB	diabetes	1=yes, -1=no	B
SHOCK	cardiogenic shock	1=yes, -1=no	B
SMOKE2	smoking history	1=never, 2=current, 3=quit <1 yr, 4=quit 1-5 yr, 5=quit5-10 yr, 6=quit >10 yr	O
NONSMOK	nonsmoker	1=yes, -1=no	B
CURSMOK	current smoker	1=yes, -1=no	B
PREVSMOK	previous smoker	1=yes, -1=no	B
HYPCHOL	hypercholesterolemia	1=yes, -1=no	B
CVA	previous cerebrovascular accident	1=yes, -1=no	B
ANEMIC	anemia	1=yes, -1=no	B
DEATH	mortality	1=yes, -1=no	B
ICUSTAY	number of days in intensive care unit (ICU)	ranges from 0 to 124 days	C
HSPSTAY	number of days in hospital after operation	ranges from 0 to 170 days	C

* C = continuous variable, B = binary variable, O = ordinal variable, N = nominal variable

As a first step to reduce the number of input variables, those variables missing more than twenty percent of the patient data were deleted (ETHNIC, MIONSET, CHF2, SMOKE2, NONSMOK, CURSMOK, PREVSMOK). The potential output variable ICUSTAY was also eliminated because it was missing data from 128 cases. From the remaining 29 input variables, several were missing information (unavailable or system missing) for just a few cases (less than 50 cases). Here, the missing patient cases rather than the entire variable were simply removed from the database (DISLMAIN, EJFRAC, DISMR, DISAS, HYPERTEN, MI, UNSTANG, PVD, HYPCHOL, CVA). HSPSTAY was also missing several cases which were deleted as well. Given that the transfer function used in these experiments will be the hyperbolic tangent, all binary variables were recoded to -1 and 1 to represent an absence of a characteristic or a presence, respectively. AGE and SURGDT were normalized by subtracting the mean and dividing by twice the standard deviation to give a zero mean and approximately unit variance.

Table 3.4 presents the remaining input parameters that were initially considered when developing the risk model, and their respective prevalence in the dataset. This left a database of 3427 complete cases, 29 input variables, and four output variables (DEATH, HSP5, HSP6, HSP7). The mortality rate of this final set was 3.7 percent, or 127 deaths.

Table 3.4: List of risk factors presented to the ANN

Type	SFHI Variable	Definition	Prevalence
INPUTS	SURGDT	date of surgery	N/A
	AGECAL	patient's age	N/A
	FEM	female gender	0.247
	OPMVD	mitral valve disease operation	0.021
	OPAVD	aortic valve disease operation	0.031
	EMERURG	emergent/urgent priority for surgery	0.229
	PTCA	failed PTCA prior to surgery	0.047
	REOP	reoperation	0.100
	RENLDX	renal disease	0.062
	DISVAN	ventricular aneurysm	0.003
	DISLMAIN	left main disease	0.195
	EJFRAC	ejection fraction: normal moderate severe	 0.660 0.250 0.090
	DISMR	mitral valve regurgitation	0.041
	DISAS	aortic valve stenosis	0.025
	DISTRIV	tricuspid valve disease	0.001
	HYPERTEN	hypertension	0.621
	PREIABP	pre-operative intraaortic balloon pump (IABP)	0.042
	MI	previous myocardial infarction (MI)	0.532
	EVOLMI	evolving MI	0.013
	HXCHF	history of congestive heart failure	0.102
	UNSTANG	unstable angina	0.629
	CVDDX	cerebrovascular disease	0.119
	PVD	peripheral vascular disease	0.121
	TVD	triple vessel disease	0.740
	OBESE	obesity	0.107
	SMALL	small stature	0.018
	COPD	chronic obstructive pulmonary disease	0.132
	DIAB	diabetes	0.228
	SHOCK	cardiogenic shock	0.007
	HYPCHOL	hypercholesterolemia	0.531
	CVA	previous cerebrovascular accident	0.039
	ANEMIC	anemia	0.165
OUTPUTS	DEATH	mortality	0.037
	HSP5	five-day stay in hospital after operation	0.136
	HSP6	six-day stay in hospital after operation	0.247
	HSP7	seven-day stay in hospital after operation	0.379

3.5 Selection of Important Variables

An important point to consider when selecting the important variables is the problem of overfitting. Many previous CABG risk models have employed a relatively large number of variables in their analysis. This can lead to overfitting the model to the original database, thereby reducing the model's ability to generalize. The general rule is to attempt to develop the best possible model with the fewest variables – a parsimonious model. As the number of variables increases, the model improves (this is inherent), however, sometimes only slightly. On the other hand, increasing the number of variables decreases the power and speed of the analysis that are also important factors [Tabachnick & Fidell 1989]. If the model has little power, then it is not useful in predicting outputs for other databases. Considering this information, reducing the number of risk factors used in the final model is advisable.

The original variables used in the model development are presented in Table 3.4. The variable set was reviewed by Shaw at the SFHI to ensure that the selected variables agree with the variable definitions, since substituting variables with data that provide similar information is sometimes possible. An example is a variable that measures serum creatinine may be replaced by the presence of renal disease [Pliam *et al.* 1997]. The next step was to employ a variable-reduction technique to reduce the size of the risk model. Two variable-reduction approaches were considered: one using statistics, and the other using the weight-elimination cost function for backpropagation ANNs.

3.5 (a) Statistical Approach

The following is an outline of the statistical approach typically employed by medical researchers to identify the most important input variables. First, univariate analysis is used to observe the relationship between each variable and the outcome of interest, namely, in-hospital death. To retain as many important factors as possible using stepwise selection, a significance level of $p = 0.20$ to enter and $p = 0.10$ to retain allows liberal inclusion of potentially important variables [Edwards *et al.* 1997]. Although univariate analysis does not take into consideration the influence of other variables while observing one in particular, it is a common first step when trying to remove inconsequential variables [Grover *et al.* 1996].

The next step is to use regression analysis to identify the most important variables from the reduced variable set. When the outcome of interest is dichotomous, logistic regression analysis is commonly used. Logistic regression is well-suited for this type of analysis, because its output ranges

from 0 to 1, an output that can easily be interpreted as a risk probability. Logistic regression can identify variables that are significant while taking into consideration the influence of the other variables in the equation. Theoretically, this technique will further reduce the set of variables, and a statistical risk model will result.

Although this is a valid method of developing a risk model for this database, the type of data (categorical) involved with a coronary artery database complicates its analysis. Therefore, this approach of statistical modelling will be put aside as future work for a graduate student in statistics. This thesis will focus on the second method of variable reduction, the weight-elimination cost function of an ANN.

3.5 (b) ANN with Weight-Elimination

The weight-elimination cost function of an ANN has the power to select the most important input variables by reducing the weights of the inconsequential variables to zero [Weigend *et al.* 1990a, Trigg 1997, Frize *et al.* 1997, Ennett & Frize 1998]. The initial input variable set should include the risk factors that other researchers have identified as important parameters using univariate analysis, multivariate analysis, and clinical intuition. Pliam *et al.* [1997] performed these analyses to find the most influential variables for the SFHI database, but did not discuss which variables were chosen by which method. It is possible that if more variables were considered, the weight-elimination cost function might find variables that had previously been considered unimportant are quite influential as Baxt [1994a] found with acute MI. Nevertheless, these newly-identified influential variables would have to show a link to the situation under consideration. Otherwise, they may just be coincidental. Therefore, the initial variable list for the experiments using the SFHI database was based on the factors identified by univariate analysis in their paper [Pliam *et al.* 1997], and on other variables that newer models, published after Pliam *et al.*'s work was presented, have selected (refer to Table 3.4).

Introducing this list of variables to the ANN with weight-elimination should hypothetically reduce the variable set to the most significant factors. A potential problem with this dataset is the prevalence of binary coded variables. Warner [1997] suggested that ANNs may require a more complex database to deduce the nonlinear relationships between the input variables. Faussett [1994] and Penny & Frost [1996] have suggested that ANNs can learn distinct categorical responses easier than a continuous variable. Nevertheless, categorizing a continuous variable will mean a loss of valuable information about cases that occur on or near the boundaries of the groups.

Also, using a double-layered network with a variable number of hidden nodes increases the complexity of the model, as well as making the interpretation more difficult. First, a double-layered network means that there are two levels of weights, one for each layer. Second, each node in the hidden layer will have its own weight for each variable. Therefore, if there are eight hidden nodes, then eight weights should be considered when deciding the overall importance of the variable. A factor which one node may consider inconsequential may be important for another node. Therefore, a variable should only be deemed unimportant if all weights on all nodes and in all layers are near zero. Weight-elimination can thus be used to reduce the number of input variables, while reducing the complexity of the network and achieving a parsimonious model.

Chapter 4: Coronary Artery Surgery Mortality Model

This chapter outlines the final network configuration, and justifies the specific architectural choices. Then, details of how the training and test sets were developed are presented. In particular, an indepth discussion of creating the artificial datasets for experimentation requires a brief description of various techniques to build these sets. Finally, the set of experiments under investigation are recapped.

4.1 The ANN Design

The literature search revealed the optimal selection for the model configuration. Based on previous experiments in other medical domains as well as the problem of CABG surgery mortality prediction, the pros and cons of many approaches have been weighed. Since experiments investigating the performance of additive and statistical models on the SFHI database were completed by Pliam *et al.* [1997], the objective of this thesis was to observe the effectiveness of an artificial neural network on the same database. Thus, the experimental simulations will be carried out with an ANN having the following configuration:

- backpropagation training algorithm
- hyperbolic tangent transfer function
- weight-elimination cost function
- double-layered architecture

Many researchers have compared the various ANN architectures, and concluded that the best-performing network was the feedforward ANN trained with the backpropagation training algorithm. The backpropagation-trained ANNs have outperformed other training methods and other modelling techniques [Buchman *et al.* 1994, Burke *et al.* 1994, Trigg 1997, Frize *et al.* 1997]. This training algorithm has also been successful in predicting outcomes in a variety of medical settings including

cardiac surgery [Baxt 1991, 1993, 1994, Lippmann & Shahian 1997]. These findings support the choice of a feedforward backpropagation ANN for running these experiments.

Because of the error term in the backpropagation learning algorithm, no learning occurs when the value of the weight is zero, because the resultant weight change would also be zero (refer to Appendix C). Using an output range of -1 to 1 with inputs scaled to have zero mean and unit variance, the network can learn faster [Penny & Frost 1996]. A nonzero output is more likely because the transition region (where the values shift from one end of the spectrum to the other) of a sigmoidal function that ranges from -1 to 1 occurs at zero [Fausett 1994]. Furthermore, given the small number of sample patterns for nonsurvivors in the SFHI database, the network requires as much training information as possible. Therefore, the network used in these experiments contains the hyperbolic transfer function with the hope that it will be better able to distinguish between patient survivors and nonsurvivors.

A backpropagation network with the weight-elimination cost function described by Weigend *et al.* [1990a, 1991a, 1991b] offers several benefits over the standard backpropagation technique. First, it acts as a variable-reduction method by forcing already small weights to zero, hence eliminating the influence of those variables on the ANN model. This reduces the complexity of the model, which could improve the model's generalizability to new data. Secondly, the largest weights on the remaining variables have been shown to affect the outcome [Weigend *et al.* 1991a, 1991b, Trigg 1997, Frize *et al.* 1997, 1998]. By reducing the input variable set to just the largest weights, the performance of the ANN was not compromised, in fact, it was enhanced [Frize *et al.* 1998]. The ANN to be employed in this work has already successfully been validated. Trigg [1997] compared the performance of this ANN using standard backpropagation and backpropagation with weight-elimination with the results of Weigend *et al.* [1990a, 1990b, 1991a, 1991b] using the Tong's [1983] sunspot dataset. The results confirmed that both algorithms are working properly [Trigg 1997, Frize *et al.* 1997]. Therefore, it is with confidence that these experiments were carried out using the weight-elimination cost function. For the standard backpropagation-trained ANN, the cost function minimized was the sum of squared errors (SSE) between the target values and the network's outputs.

Recall from Chapter 2, many researchers have highlighted the benefit of using multilayered networks over single-layered networks [Hornik *et al.* 1989, Fausett 1994, Penny & Frost 1996, Trigg 1997, Frize *et al.* 1997]. Having more than one layer allows the ANN to develop models with nonlinear characteristics, if necessary. This is not possible with a single-layered ANN. Despite the advantages of

multilayers, most researchers agree that although more than one hidden layer is advantageous in some circumstances, for the most part one hidden layer is sufficient. More hidden layers also significantly increase the complexity of the model because one must also take into consideration the additional hidden nodes involved. Therefore, to facilitate the results' interpretation while maintaining a certain level of computational power, the optimal choice is a double-layered construction.

The feedforward backpropagation ANN was originally programmed by Buskard [1994] and modified by Trigg [1997] for use in the matrix program, MATLAB®. This tool incorporates several useful functions that allow users to modify the experiments. It was possible to use either the standard backpropagation algorithm or backpropagation with the weight-elimination cost function. The user can choose between single- and double-layered architectures. The number of hidden nodes in the double-layered ANN is also adjustable. Modules are also included within the program that calculate the *a priori* statistics for the training and test sets, as well as the operating points required for constructing an ROC curve.

Trigg [1997] modified the standard backpropagation architecture's routines where the error and weight updates occur to integrate the weight-elimination cost function developed by Weigend *et al.* [1990a, 1991a, 1991b]. The particular MATLAB® Neural Network Toolbox files that were adapted to accommodate the weight-elimination technique were *sumsqr.m* and *learnbpm.m*. The weight-elimination function can be toggled on and off using a switch in the main program. When the weight-elimination routine is turned off, the ANN uses the SSE as its error function [Trigg 1997].

4.2 Generation of Training and Test Sets

The SFHI reduced database was separated into a training set and a test set to develop the ANN model. This approach is called cross-validation. For the most part, the SFHI database preprocessing was performed using the statistical program SPSS®. Using an SPSS® function that randomly creates subsets of the database, the SFHI data was divided into a training set (the cases upon which the models will be developed), and a test set (the cases that will be used to test the performance of the models) containing two-thirds and one-third of the cases, respectively. This is a common approach to separate the database when the number of sample cases for the underrepresented class is small or when using a small dataset. The resulting training set contained 2254 cases with a 3.7 percent in-hospital mortality rate (83 cases of

nonsurvivors), while 44 of the 1173 patients in the test set died in-hospital postoperatively (giving a mortality rate of 3.8 percent).

Recall that a difficult obstacle to overcome with coronary artery surgery databases, from a researcher's point of view, is the low mortality rate. This means that there are not many sample cases to represent the characteristics of a patient who will not survive the heart surgery. Having few sample cases makes it difficult to identify the risk factors for operative death. This thesis involves an investigation of an innovative technique for dealing with a small sample of data of nonsurvivors: modifying the case distribution.

4.2 (a) Modifying the Case Distribution

Several techniques are available to potentially improve the risk model's performance. The objective here is to artificially increase the percentage of nonsurvivors to more easily identify their characteristics.

Since Baxt & Skora [1996] showed that an ANN trained on a database with a higher-than-reality prevalence performed better than one based on the actual disease-state prevalence, the duplication method to increase the occurrence of nonsurivors in the training set is appealing. To prevent the loss of important information about the survivors, duplicating the nonsurvivors (and essentially increasing the mortality rate by artificial means) may be a better option.

The next step is to choose the method of preparing the artificial datasets. Two commonly used techniques to develop simulated datasets, the jack-knife method and the bootstrap approach, are described briefly in the next sections of the chapter, followed by a description of a technique developed by Katz *et al.* [1994] that combines several approaches.

4.2 (a) i. Jack-Knife Method

Medical applications often do not have enough data to develop sufficiently large training and test sets, and rarely enough for a separate validation set to evaluate the model. One solution to this problem is the jack-knife method. This technique trains n different sets built from n cases while leaving out one case each time to test the network and assess the learning. This "leave-one-out" approach can provide a good estimate of the network's performance, but it is time consuming because n different ANNs must be trained [Forsström & Dalton 1995].

An "adapted jack-knife" method or a "multiple holdout" sample technique reduces the number of training sets required. In this situation, the database is randomly divided into a small number of subsets, say three or five, and each time the network is trained while one subset is withheld for testing. This process is repeated numerous times (30-100 repetitions are acceptable to be statistically stable), and the network performance is averaged over all test sets [Forsström & Dalton 1995, Kattan & Beck 1995].

4.2 (a) ii. Bootstrap Method

The bootstrap method is less computationally intense than the jack-knife approach. The database can be randomly separated into a training and test set (the training set should be larger, say, two-thirds/one-third). Based on the assumption that each case in the test set is unique (i.e., each case has a probability of occurrence of $1/n$), a large number (about 100) artificial test sets are created by randomly selecting with replacement n cases from the original test data. Random selection with replacement indicates that one case may be entered into a particular dataset more than once or not at all, therefore the test sets are different. For this method, the network is trained once, and tested several times [Forsström & Dalton 1995].

4.2 (a) iii. Neural Net-Bootstrap Technique

Katz *et al.* [1994] developed a hybrid risk modelling approach to estimate the occurrence of postoperative valve-related deaths of patients implanted with artificial heart valves. Katz *et al.* [1994] used an artificial neural network that applied a bootstrap sampling technique to measure prediction accuracy and for estimating prediction errors. To account for the low mortality rate (60/776), they created artificial training and test datasets by separating the patient records into two groups (survivors and nonsurvivors), and then randomly sampling with replacement. Then, the ANN learned the sample patterns from the training set, and the test set estimated the efficacy, determined the prediction errors, and was used to optimize the network. The system could then select new training and test sets according to the previous procedure, and the process continued. The procedure was executed 1300 times, and the test results were averaged over all test sets. The results showed that the patients were correctly classified by the network 78 percent of the time using only preoperative information with an averaged error of prediction of 22 percent ± 5 percent [Katz *et al.* 1994].

4.2 (b) Experimental Datasets

The objective was to artificially increase the number of nonsurvivors (and hence the mortality rate) in the training and test sets. To achieve this goal, the datasets were separated according to their

outcome: death or survival. This approach was necessary due to the small number of nonsurvivors, and to ensure that the desired distribution could be achieved. By combining the idea of separating the patient records according to the outcome used by Katz *et al.* [1994] with the bootstrap sampling technique so that a specific mortality rate could be obtained, the artificial datasets for the SFHI coronary artery surgery database were created. Therefore, once the original database was divided into training and test sets (two-thirds and one-third, respectively), these sets were further subdivided into those who survived the surgery and those who did not. The artificial sets were formulated from these "categorized" datasets.

In addition to the two datasets with the true mortality distributions, four different artificial datasets were developed: three artificial training sets, and one artificial test set. First, there were the datasets with the true mortality rate distribution. To artificially increase the percentage of nonsurvivors in the datasets, a simple program that performed random sampling with replacement was used. Since the sample size can affect the performance of a model, the total number of cases in the training and test sets was kept constant. Given the number of cases, and the desired percentage of nonsurvivors for the particular dataset, it was possible to determine how many patients who did not survive the surgery would be included in that particular set. The three artificial training sets had mortality rates of 10, 20 and 30 percent, respectively. This offered a range of training set mortality distributions for the experiments to be performed. Table 4.1 describes each dataset with the number of nonsurvivor and survivor patient cases. The nonsurvivor and survivor cases were chosen separately, but in each situation, the patient records were chosen randomly with replacement. The artificial test set was designed to have a 20 percent mortality rate, and Table 4.2 describes the distribution between the survivors and nonsurvivors for these test sets. In addition to these artificial datasets, another 30 different artificial test sets with a mortality rate of 20 percent and 30 sets with the true distribution were created. These additional test sets were used for the bootstrap approach to provide a number of datasets upon which to test the ANN. The results were averaged over the 31 test sets.

Table 4.1: Distributions of the training datasets

Dataset	# Cases nonsurvivors	# Cases survivors	Total
3.7 % true distribution (trorig)	83	2171	2254
10% mortality rate (tr10)	225	2029	2254
20% mortality rate (tr20)	451	1803	2254
30% mortality rate (tr30)	676	1578	2254

Table 4.2: Distribution of the test datasets

Dataset	# Cases nonsurvivors	# Cases survivors	Total
3.8 % true distribution (teorig)	44	1129	1173
20% mortality rate (te20)	226	947	1173

Initially, the plan was to observe the effect of using artificial training and test sets in an attempt to improve the ANN performance. The original hypothesis was that training a network on a higher-than-normal prevalence would improve the prediction ability of the risk model. The second initiative was to also develop artificial test sets, and observe the effect on the performance of a network trained using various training sets (artificial and natural), and then tested on different test sets (artificial and natural). Due to the large amount of time required to optimize the best-performing networks (choosing how many nodes, tuning the algorithm parameters like learning rate, momentum, etc.), focussing on fewer experiments was essential. Providing fewer experimental results that are more detailed is better than performing many simulations and not being able to recognize the important aspects of the networks.

To summarize, the experimental datasets included four training sets (one with the true distribution, and three artificial datasets with the following mortality rates: 10, 20, 30 percent) and two test sets (one that was the true distribution, the other with a mortality rate of 20 percent). The final experimental simulations involved a focus on three aspects of ANN performance:

1. Observe the effect of training an ANN with one set of higher-than-normal prevalence and one set with the true distribution (theoretically, the same mortality rate as the original test set), and test it on the true distribution of the test set (Baxt's approach).
2. Observe the effect of training an ANN on three training sets with different mortality rates (higher and lower mortality rates than the test set) and assessing its performance when an artificial test set with a 20 percent mortality rate is used to validate the model.
3. Take the overall best-performing network, and investigate its connection weight values. Eliminate the variables whose connection weights are zero, and retrain the network using only the remaining weights to see the effect on the ANN performance.

Chapter 5: Simulation Results and Evaluation

This chapter presents the results of the ANN experimental simulations using the SFHI coronary artery surgery database. The calibration of the ANN's settings is first considered to achieve the network's optimal level of performance. Once the parameters have been fine-tuned to maximize the performance of each ANN for the five situations under investigation (trorig/teorig, tr20/teorig, tr10/te20, tr20/te20, tr30/te20), the networks are run using weight-elimination, and then rerun under the same settings without weight-elimination. The results are evaluated based upon the sensitivity of the networks. A model with higher sensitivity means that more of the difficult, high-risk patients (the patients who do not survive the surgery) are correctly identified than the other models. The results of these experiments are compared with Trigg's experiments on an adult ICU database. A comparison was also made with the SFHI research group's work, which apply different additive and statistical risk models. A discussion of the problems encountered when working with highly skewed datasets follows. The chapter rounds out with some general comments about ANNs applied to the coronary artery surgery medical environment.

5.1 Network Calibration

As described in Chapter 4, the ANN program has several adjustable parameters that modify the performance of the network. The random number generator (RNG) seed that selects the initial network weights was 18 (this was the value chosen for the double-layered networks by Trigg [1997], however, any number could have been used). The initial weights for the input layer were randomly chosen to be between -1 and 1 using the random generator function in MATLAB®, whereas the hidden layer weights were between -0.1 and 0.1. These relatively small values were used by Trigg [1997] and recommended by Weigend *et al.* [1990a, 1990b] to achieve improved performance with backpropagation networks. The preliminary simulations required many trials to explore the effects of tuning the learning rate, adaptive

learning rate features, weight-elimination constant, weight-elimination scale factor, momentum, error ratio, the number of nodes in the hidden layer, the output error weighting factor, and the cutoff value. Table 5.1 summarizes the adjusted parameters and the approximate range over which the ANNs were tested. From these first experiments, reasonable ranges were identified to achieve "acceptable" performance. Then these parameters were fine-tuned for each network to find the optimal settings.

Table 5.1: Optimization parameters and their approximate range implemented

Parameter	Range
Learning rate (lr)	0.0005 → 0.01
Learning rate increment (lr_inc)	1 → 1.05
Learning rate decrement (lr_dec)	1 → 0.95
Weight-elimination constant (λ)	0.00003 → 0.005
Weight-elimination scale (w_0)	0.1 → 0.3
Momentum (α)	0 → 0.99
Error ratio (err_ratio)	1.001 → 1.05
Hidden nodes	1 → 10
Output error weighting factor (wtfact)	1 → 1.50
Cutoff value (float)	-0.20 → 0.20

5.2 Performance Measure Optimization

Given the modelling challenges presented by a CABG database with a low mortality rate outlined in Chapter 4, it was necessary to re-evaluate how the ANN model's performance should be measured. The following sections discuss the reasons behind the choice of sensitivity as the best measure of the network's classification performance, why the reported results are based on the test set, and show the sensitivity of the training set to the initial random weights.

5.2 (a) Sensitivity as the Measure of Best Performance

Due to the highly skewed distribution towards the survivors in this CABG database, the CCR and ROC curves are not the best measures of performance in this situation, as discussed in Chapter 2. Therefore, the selection of the best-performing networks focussed on the sensitivity of the test set, in other words, the CCR of the actual nonsurvivors. The focus for these experiments was the classification of the nonsurvivors, since they are the most difficult patients to identify. It is important to remember that these mortality risk models are not yet accurate enough to be used as clinical tools, and a patient should not be denied surgery simply because the ANN's output predicts death.

Although sensitivity was considered the deciding factor in selecting the best networks, there were several other factors that influenced the choice of the optimal network. The best-performing network was selected based on a balance of the criteria outlined below. The criteria deemed important for choosing the best-performing ANN (from most important to least important) were:

- highest sensitivity for the test set;
- highest specificity for the test set; and
- highest CCR for test set.

It is important to note that even though having a high sensitivity and a high specificity will result in a high CCR, it may not be the highest achieved by the network. If the ANN were to classify everything as belonging to the highest *a priori* probability (like a constant predictor does), the CCR would be 96.2 percent, which would be remarkably higher than a network with higher sensitivity values.

5.2 (b) Optimal Performance of the Test Set

The selection of optimal performance was based on the test set. While this is a common practice in modelling approaches, it is not always the best approach. The original plan was to choose the best-performing network given its performance on the training set, since this is a more unbiased approach. When this strategy was implemented, an interesting pattern became evident. Comparing the point at which the ANN performed best on the training set and the test set, it became obvious that the training set's maximum sensitivity occurred when the same or more nodes (compared with the test set) were included in the hidden layer. The test set's best models generally had fewer nodes than those of the training set. There was a concern that when there were more nodes, the network had begun memorizing the training set patterns, as evidenced by the diminishing classification performance of the nonsurvivors in the test set. Table 5.2 shows at what number of nodes the different networks achieved their best performance for both the training and test sets, and the sensitivity of both sets at those points.

Table 5.2: Number of nodes to achieve best-performing weight-elimination networks

Experiment	Nodes in training set at point of optimal performance	Sensitivity (train) %	Sensitivity (test) %	Nodes in test set at point of optimal performance	Sensitivity (train) %	Sensitivity (test) %
trorig/teorig	10	33.7	9.1	9	26.5	11.4
tr20/teorig	7	84.5	34.1	7	80.0	43.2
tr10/te20	8	52.7	14.4	7	46.7	16.8
tr20/te20	5	81.4	21.7	2	61.6	51.8
tr30/te20	7	91.0	23.9	2	83.0	59.3

Although Hornik *et al.* [1989], Fausett [1994], and Penny & Frost [1996] stated that a double-layered ANN can model any function given that the function and its derivative are continuous, and an adequate number of hidden units, this does not necessarily mean that more nodes are better when

concerned with generalization. Certainly, it is possible that an ANN with the above characteristics may model any function, however, such a model may be too closely related to the information in the training set, and not generalizable to new cases. Therefore, although it appears to go against the guidelines set out by the above researchers, even having as few as two hidden nodes does not necessarily hinder the performance of the ANN as shown in the experiments performed in this thesis work.

5.2 (c) Sensitivity to Initial Random Weights

Once the ANN was optimized for the training set with the particular initial random weights, preliminary results showed that changing the RNG seed caused dramatic changes in the network's classification ability. For example, consider Figure 5.1 that shows the effect of changing the RNG seed on the training set. This figure clearly shows how sensitive the network is to the initial random weights for tr20/te20. Therefore, once the ANN was optimized for the set of initial weights, the evaluation of the network focussed on the performance of 30 randomly selected test sets.

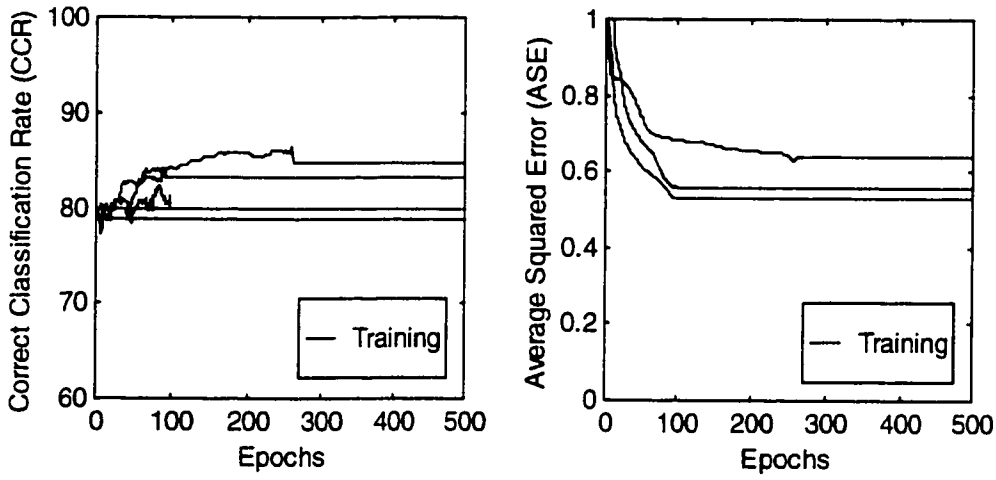


Figure 5.1: Sensitivity of the training set to the initial random weights for an ANN with weight-elimination trained on an artificial dataset with a 20 percent mortality rate and tested on an artificial test set with a 20 percent mortality rate (tr20/te20)

5.2 (d) Weight-Elimination Versus No Weight-Elimination Networks

Although the original objectives of this thesis were not to compare weight-elimination networks with those without weight-elimination, after some thought, this comparison was considered interesting and necessary to justify the use of the weight-elimination cost function. Therefore, once the weight-elimination ANNs were optimized, the networks were retrained on the same parameter settings with the standard backpropagation algorithm using the sum of squared errors as a cost function. This way, the networks trained using standard backpropagation ANNs could be used as a baseline comparison for evaluating the performance of those employing the weight-elimination technique. The analysis of these results is presented in Section 5.5 (c).

5.3 Parameter Optimization

Preliminary results showed that certain approaches were necessary to elicit the best ANN model using the CABG database. This section presents how the weight-elimination constant was used, the selection of the cutpoint value, and why single-layered networks were not developed for this database. Of course, each point will be justified with evidence to back up these choices.

5.3 (a) Weight-Elimination Constant

Although Weigend *et al.* [1990a, 1990b] recommend starting all networks with a weight-elimination constant, λ, equal to zero, and then gradually increasing its value, this technique did not always work well with this coronary artery surgery database. The objective of Weigend's approach was to let the ANN take the information from the interaction of all of the input variables first, and then start eliminating less important variables. As shown in Figure 5.2, preliminary results indicated that this procedure resulted in improved nonsurvivor classification for the training set (higher sensitivity) and poorer performance for the test set (lower sensitivity); in other words, overfitting. Therefore, in some cases, the experiments were conducted with a stationary value of λ that was near zero, but still offered better generalizability for the test set than an initial zero value that is updated after several hundred epochs.

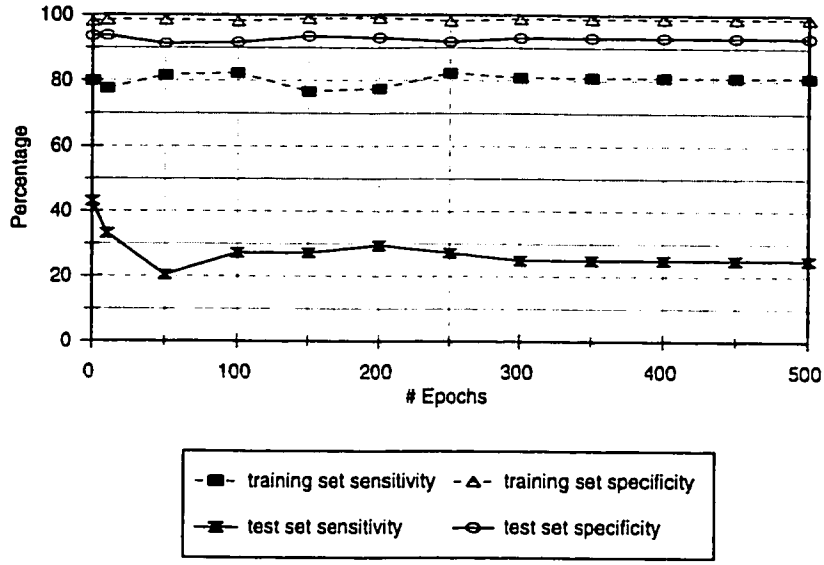


Figure 5.2: Effect of gear-shifting the weight-elimination constant of 0.0003 after *x* epochs for an ANN trained on an artificial dataset with a 20 percent mortality rate and tested on the true mortality distribution (tr20/teorig)

5.3 (b) Cutoff Value

As shown in Table 5.1, the float, or cut-point value, was one of the parameters that was varied to observe its effect on the ANN performance. In the end, it was decided to leave the cut-point at zero (since the output varies from -1 to 1, and zero is the decision boundary for this range). The reason for this decision was that although decreasing the cut-point to favour the correct classification of more nonsurvivors, the misclassification rate of the survivors increased (refer to Figures 5.3 and 5.4 that exemplify this fact using the tr20/te20 database). As shown in Figure 5.3, the operating point for the training set (tr20) (the point at which there is the best balance between the sensitivity and the specificity) is at the cutoff value of -0.05 on a scale from -1 to 1. The operating point of the test set (te20) is at a lower cut off value of -0.45. Comparing the CCR of the test set for cut off values of -0.45 and 0 in Figure 5.4, it is evident that the CCR for the cut off of -0.45 is about 8 percent lower than for the 0 value. Therefore, once again taking into consideration the balance between the sensitivity, specificity and CCR, the correct classification of an additional 37 nonsurvivors at the expense of the misclassification of an additional 82 survivors was not considered reasonable. This ratio of classification and misclassification of nonsurvivors and survivors is even more prominent in the cases where the network was trained on a higher prevalence.

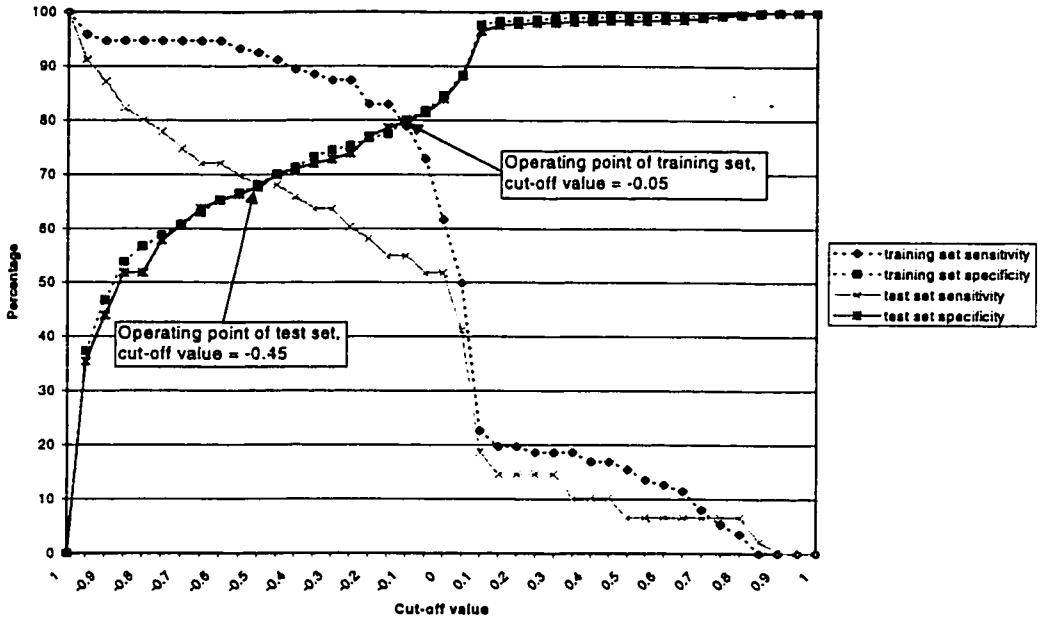


Figure 5.3: Sensitivity and specificity of the training and test sets for an ANN with weight-elimination trained on an artificial dataset with a 20 percent mortality rate and tested on an artificial test set with a 20 percent mortality rate (tr20/te20)

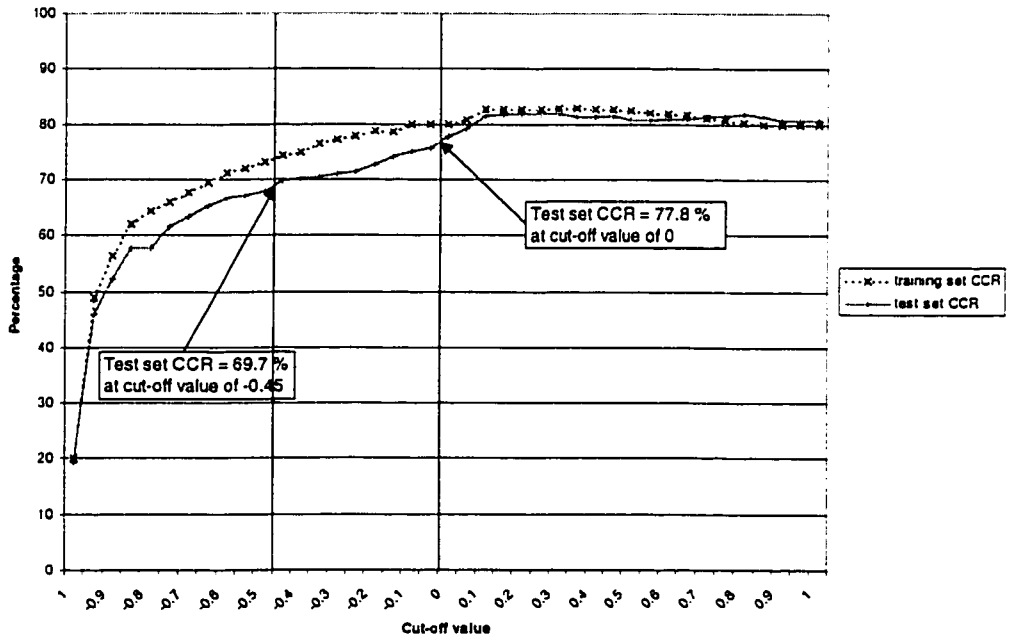


Figure 5.4: Correct classification rate of training and test sets for an ANN with weight-elimination trained on an artificial dataset with a 20 percent mortality rate and tested on an artificial test set with a 20 percent mortality rate (tr20/te20)

5.3 (c) Single- Versus Double-Layered Networks

One last consideration was the number of layers in the ANNs. Although the advantages of double-layered networks were discussed in Chapter 4, it was still necessary to test the network performance of a single-layered network on this CABG database. The results shown in Table 5.3 of the preliminary testing of the single-layered networks indicate that they were unable to improve the classification rate of the nonsurvivors (i.e., single-layered networks had lower sensitivity on the test set than double-layered ANNs).

Table 5.3: Comparison of single- versus double-layered network performance for an ANN with weight-elimination trained on an artificial dataset with a 20 percent mortality rate and tested on an artificial test set with a 20 percent mortality rate (tr20/te20)

No. of layers	Training set sensitivity (%)	Training set specificity (%)	Training set CCR (%)	Test set sensitivity (%)	Test set specificity (%)	Test set CCR (%)
1	58.1 (262/451)	86.4 (1558/1803)	80.7	38.5 (87/226)	87.9 (832/947)	78.3
2	61.6 (278/451)	84.6 (1525/1803)	80	51.8 (117/226)	83.9 (795/947)	77.7

The results in Table 5.3 justify the use of the double-layered ANN architecture with the SFHI coronary surgery database. In retrospect, given the complexity of the CABG patient database (i.e., the fact that the database is quite homogeneous), it seems reasonable to accept that this particular database would require a more complex architecture to optimize the network's performance.

5.4 Results

Given the information in the above sections, the best-performing weight-elimination networks were selected based on those criteria, and Tables 5.4 and 5.5 summarize the parameter settings and performance of these networks. Table 5.4 provides the parameter settings at which the optimal performance was achieved: number of layers, RNG seed, initial weights, number of hidden nodes, learning rate and its adjustable parameters, weight-elimination constant, weight-elimination scale factor, momentum, error ratio, error weighting factor, and the cut off value. Table 5.5 provides a summary of the performance measures based on the test set performance: sensitivity, specificity, PPV, PNV, CCR, and the C-index (as defined in Section 2.2, pages 8-11). Note that the CCR of the constant predictor is also included in this table, so it could be used for comparisons in later sections. The CCR curves with the ASE and the ROC curves are available for each training set/test set combination for reference in Appendices E and F, respectively. The results are the average (mean) performance of the 31 different

test sets for each experiment, and the standard deviation of those experiments. All test sets within each group (teorig and te20) are the same size with the same prevalence. The results from these individual test sets are available in Appendix G.

Table 5.4: Architectures of the best-performing double-layered weight-elimination networks

Experiment	trorig/teorig	tr20/teorig	tr10/te20	tr20/te20	tr30/te20
No. of layers	2	2	2	2	2
RNG seed	18	18	18	18	18
Initial weights	W1=rands()*1 B1=rands()*1 W2=rands()*0.1 B2=rands()*0.1	W1=rands()*1 B1=rands()*1 W2=rands()*0.1 B2=rands()*0.1	W1=rands()*1 B1=rands()*1 W2=rands()*0.1 B2=rands()*0.1	W1=rands()*1 B1=rands()*1 W2=rands()*0.1 B2=rands()*0.1	W1=rands()*1 B1=rands()*1 W2=rands()*0.1 B2=rands()*0.1
Hidden nodes	9	7	7	2	2
lr	0.001	0.001	0.001	0.001	0.0001
lr_inc	1.003	1.003	1	1.003	1.003
lr_dec	1	1	1	1	1
λ	0.0003 after 0 epochs	0.0001 after 0 epochs	0.0004 after 10 epochs	0.0003 after 10 epochs	0.0003 after 0 epochs
w_0	0.15	0.1	0.1	0.1	0.1
Momentum	0.75	0.88	0.1	0.5	0.35
err_ratio	1.02	1.02	1.02	1.02	1.02
wtfact	1	1	1.4	1.2	1
Float	0	0	0	0	0

Table 5.5: Performance measures for the best-performing double-layered weight-elimination networks averaged over the 31 different test sets (mean ± standard deviation)[6]

Experiment	trorig/teorig	tr20/teorig	tr10/te20	tr20/te20	tr30/te20
Sensitivity (%)	10.85 ± 4.38	43.55 ± 7.59	18.38 ± 2.27	53.23 ± 2.76	61.68 ± 11.85
Specificity (%)	98.26 ± 0.39	93.83 ± 0.77	96.35 ± 0.51	82.86 ± 1.06	79.82 ± 0.82
PPV (%)	19.42 ± 6.93	21.64 ± 3.46	54.62 ± 4.50	42.59 ± 2.04	43.02 ± 1.30
PNV (%)	96.59 ± 0.16	97.71 ± 0.30	83.19 ± 0.39	88.13 ± 0.64	90.25 ± 0.75
CCR (%)	94.98 ± 0.36	91.94 ± 0.79	81.33 ± 0.58	77.15 ± 1.04	76.74 ± 0.74
C-index	0.9387 ± 0.0029	0.8978 ± 0.0070	0.7809 ± 0.0042	0.7285 ± 0.0052	0.7535 ± 0.0057
CCR of constant predictor (%)	96.2	96.2	80	80	80

5.5 System Evaluation

This section comprises the separate evaluation of each of the comparable sets of experiments: modifying the training set, training with different training set prevalences, and the use of weight-

[6] The results presented here are the means and standard deviations over 31 test sets derived from an original test set. All test sets within each group (teorig and te20) are the same size with the same prevalence. The results from these individual test sets are available in Appendix G.

elimination. Each evaluation will consist of comparisons of the individual performance measurement parameters (sensitivity, specificity, PPV, PNV, CCR, and C-index), followed by a discussion of which training set/test set combination was the best. The results are also compared with the CCR of the constant predictor that offers some insight into the performance value of these networks.

5.5 (a) Evaluation of Modified Training Sets Approach

This subsection involves a comparison of the ANN performance on the true distribution test sets when training with the true distribution (i.e., a mortality rate similar to the test sets) or an artificial training set with a 20 percent mortality rate. In other words, this is a comparison of the performance of trorig/teorig and tr20/teorig. The question here is: does training with a higher-than-normal prevalence improve the ANN performance of the test sets with the true mortality rate?

Looking at the performance measures in Table 5.5 individually, it is clear that training with a higher-than-normal prevalence has a dramatic effect on the mean sensitivity rate. In the case of trorig/teorig, the mean sensitivity of the test sets was 10.85 percent, but when the network was trained with the artificial dataset with the higher mortality rate, the mean test set sensitivity increased to 43.55 percent — a remarkable improvement of nearly 33 percent. This result indicates that there was a noticeably larger number of nonsurvivors correctly classified as nonsurvivors by the ANN. Training with a higher-than-true prevalence caused a slight reduction in mean test set specificity (a drop from 98.26 to 93.83 percent). This approach resulted in an approximately 5 percent decrease in the mean specificity of the test sets.

Table 5.5 shows that the PPV and PNV were barely affected by the change in training set mortality rate (19.42 to 21.64 percent difference for the mean PPV, and 96.59 to 97.71 percent difference for the mean PNV for training with the true mortality rate compared with the artificial mortality rate, respectively). Besides that, training with the 20 percent mortality rate resulted in a lower mean CCR and mean area under the ROC curve due to a greater number of incorrectly classified patients.

The CCR of a constant predictor for this set of simulations is 96.2 percent, given that the dominant class is the survivors and 96.2 percent of the test set patients were survivors. Although the mean CCR of the trorig/teorig combination approaches this CCR (94.98 percent), the number of nonsurvivors that were correctly classified is lower than that of the tr20/teorig experiments. Since the

nonsurvivors are the ones that challenge the network for correct classification, the trorig/teorig network is of less value than an ANN that has a higher sensitivity for the nonsurvivors.

Based on the above information, it is possible to conclude that, yes, training with a higher-than-normal prevalence of the under-represented outcome (here, in-hospital death) improves the mean sensitivity of the ANN using the test data without dramatically affecting other aspects of the network's performance. The ANN trained with a higher mortality rate also had a slight reduction in mean specificity with little effect on the PPV or the PNV. The network trained under these conditions maintained a high CCR and C-index, as well. Even though the network trained with the higher prevalence of mortality has a lower CCR than the true mortality rate, and although both networks (trorig/teorig and tr20/teorig) have lower mean CCRs than the constant predictor, the ANN trained on the higher prevalence is of more clinical value because more nonsurvivors are correctly classified. The higher risk patients are always more difficult to classify using risk stratification models, therefore, preference should be given to a model that is better able to classify these individuals [Clark 1996].

5.5 (b) Evaluation of Artificial Test Sets

The objective of these simulations was to identify the effect on the ANN's classification rate of artificial test sets with a 20 percent mortality rate when the network was trained using artificial training sets with lower, the same, and higher prevalences than the test sets (i.e., trained with mortality rates of 10, 20 and 30 percent, respectively). This involves a comparison of the following training and testing set combinations: tr10/te20, tr20/te20, and tr30/te20.

The following trends were noticeable based on the results of the ANN's average performance on the test sets as the training set CABG mortality rate increased (from 10 to 20 to 30 percent) while the test set mortality rates remained constant:

- The mean sensitivity increased (from 18.38 to 53.23 to 61.68 percent).
- The mean specificity decreased (from 96.35 to 82.86 to 79.82 percent).
- The mean PPV decreased (from 54.62 to 42.59 for the tr10/te20 and tr20/te20 combinations, however, the mean PPV for tr30/te20 was slightly higher than that of the tr20/te20 sets — its PPV was 43.02 percent).
- The mean PNV increased (from 83.19 to 88.13 to 90.25 percent).
- The mean CCR decreased (from 81.33 to 77.15 to 76.74 percent).

- The mean C-index decreased (from 0.7809 to 0.7285 for the tr10/te20 and tr20/te20 experimental combinations, however, the mean C-index for tr30/te20 was slightly higher than that of the tr20/te20 sets — its area under the ROC curve was 0.7535).

The above trends seem to be reliable because most of them also occurred with the ANNs tested on the true mortality test set. One exception is the PPV results. These results are consistent between neither the two groups of experiments (testing using teorig or te20) nor within the sets of experiments using the artificial test sets (te20).

It is curious that the mean C-index for tr30/te20 is slightly greater than that of tr20/te20 (0.7535 versus 0.7285). This pattern does not follow the proposed trends set out above. A possible explanation is that since the two experiments have similar classification measures, the slight increase in area under the ROC curve may be attributed to the smaller increase in sensitivity and small decrease in specificity.

Another interesting detail to highlight is the greater difference between the results of tr10/te20 and tr20/te20 as compared with the test set results between tr20/te20 and tr30/te20 (i.e., for the sensitivity, there was a 35.0 percent difference between the networks trained with tr10 and tr20, while there was only a 8.5 percent difference between the ANNs trained with tr20 and tr30). A possible reason is a lack of data on nonsurvivors in the training set with 10 percent mortality rate that may account for the poorer ANN test set classification of the nonsurvivors when trained with this set. Once again, although the mean CCR and mean C-indices are higher when the network is trained with the artificial set with 10 percent mortality, it is not necessarily a better model because fewer of the "important" cases (i.e., the nonsurvivors) are correctly classified. Perhaps these results also indicate that there may be a limit to which the mortality rate of the training set has an effect on the test set classification performance.

For these experiments, the CCR of a constant predictor is 80 percent, meaning that if all of the patients were categorized according to the output class with the highest *a priori* probability, the constant predictor would correctly classify 80 percent of the patients. It is clear that lower prevalence training sets achieve a CCR that is closer to the test set mortality rates, but this is not indicative of the clinical value of the model. When the training set mortality is lower, fewer of the nonsurvivors who are more difficult to classify are correctly categorized.

These experiments confirm the results found in the previous section that training with a higher prevalence than found in the test sets increases the ANN's mean sensitivity without having a drastic effect on the other performance parameters. However, it is also clear that training with a lower prevalence than that of the test sets decreases the mean test set sensitivity rate. Therefore, although tr30/te20 has the lowest mean CCR (and its CCR is lower than a constant predictor), it would be more useful in a clinical setting than the other models to help identify the patients who would not survive CABG surgery.

5.5 (c) Evaluation of Weight-Elimination Technique

As stated in Section 5.2, the weight-elimination networks were compared with their no weight-elimination counterparts that used the sum of squared errors cost function, and the results of this analysis are presented in Table 5.6. Recall that the objective of the weight-elimination cost function is to reduce the size of the connection weights to eliminate less useful input variables. Theoretically, this course of action of reducing the number of weights in the network, and hence the network's complexity, is expected to improve the network's classification performance [Weigend *et al.* 1990a, 1990b, Trigg 1997, Frize *et al.* 1997, Frize *et al.* 1998]. Of course, the question here is: does weight-elimination actually improve ANN classification performance with this particular database?

Table 5.6: Comparison of ANNs with and without weight-elimination using the best-performing networks[7]

Experiment	trorig/teorig WE*	trorig/teorig no WE	tr20/teorig WE	tr20/teorig no WE
Sensitivity (%)	10.85 ± 4.38	8.94 ± 4.54	43.55 ± 7.59	24.05 ± 7.89
Specificity (%)	98.26 ± 0.39	99.31 ± 0.23	93.83 ± 0.77	93.22 ± 0.67
PPV (%)	19.42 ± 6.93	33.15 ± 13.14	21.64 ± 3.46	12.07 ± 3.76
PNV (%)	96.59 ± 0.16	96.55 ± 0.17	97.71 ± 0.30	96.92 ± 0.31
CCR (%)	94.98 ± 0.36	95.92 ± 0.27	91.94 ± 0.79	90.63 ± 0.69
C-index	0.9387 ± 0.0029	0.9453 ± 0.0023	0.8978 ± 0.0070	0.8844 ± 0.0069

Experiment	tr10/te20 WE	tr10/te20 no WE	tr20/te20 WE	tr20/te20 no WE	tr30/te20 WE	tr30/te20 no WE
Sensitivity (%)	18.38 ± 2.27	18.38 ± 2.23	53.23 ± 2.76	42.31 ± 2.48	61.68 ± 11.85	59.26 ± 3.30
Specificity (%)	96.35 ± 0.51	96.23 ± 0.48	82.86 ± 1.06	85.57 ± 1.21	79.82 ± 0.82	82.79 ± 1.01
PPV (%)	54.62 ± 4.50	53.75 ± 3.99	42.59 ± 2.04	41.24 ± 2.44	43.02 ± 1.30	45.13 ± 2.03
PNV (%)	83.19 ± 0.39	83.17 ± 0.38	88.13 ± 0.64	86.14 ± 0.52	90.25 ± 0.75	89.50 ± 0.78
CCR (%)	81.33 ± 0.58	81.23 ± 0.52	77.15 ± 1.04	77.24 ± 1.03	76.74 ± 0.74	78.26 ± 1.05
C-index	0.7809 ± 0.0042	0.7816 ± 0.0039	0.7285 ± 0.0052	0.7384 ± 0.0050	0.7535 ± 0.0057	0.7687 ± 0.0057

* WE = weight-elimination

[7] The results presented here are the means and standard deviations over 31 test sets derived from an original test set. All test sets within each group (teorig and te20) are the same size with the same prevalence. The results from these individual test sets are available in Appendix G.

First, it is necessary to take a look at the different experiments with and without weight-elimination separately, then consolidate the findings to make generalizations, if possible. The first set of simulations involve training and testing the ANN with the true in-hospital mortality rates (trorig/teorig). In this case, weight-elimination provides a slight improvement in the mean test set sensitivity rate (1.91 percent higher) compared with the experiments without weight-elimination. The mean specificity, PNV, CCR and C-indices remain essentially unchanged. Unfortunately, the mean PPV suffers a large decrease in classification performance (a decrease of 13 percent PPV rate for the weight-elimination results), which is caused by a greater number of survivors being classified as nonsurvivors as compared with the number of correctly classified nonsurvivors.

When the network is trained with a 20 percent mortality rate and tested using the true mortality rate (tr20/teorig), the mean sensitivity increases by 19.5 percent. Again, the specificity, PNV, CCR, and C-indices remained relatively unchanged. In this situation, however, the mean PPV actually increased by 9.5 percent, a result that appears to contradict prior trends. This difference is likely related to the information contained in the artificial training set.

The first of the next set of results where the ANN trained on a dataset with a lower prevalence than the test sets (tr10/te20) showed that all of the average performance measures remained approximately the same. In this case, weight-elimination offered no improvement over the standard backpropagation training algorithm. Possibly, this may have occurred because there was not enough information in the training set to improve the network performance with weight-elimination. In other words, the training dataset may not have been sufficiently rich to offer information about the differences between the nonsurvivors and the survivors that could lead to the elimination or weight-reduction of certain input variables. Ultimately, this result indicates that the weights of the input variables were not dramatically affected by the weight-elimination parameters when training with a lower prevalence than the test sets.

When training and testing on artificial databases that have the same ratio of CABG surgery survival and in-hospital death (tr20/te20), weight-elimination still managed to effect a 10 percent increase in mean test set sensitivity over the no weight-elimination network. The other performance measures remained the same regardless of the effect of the weight-elimination cost function. Therefore, under these circumstances, there was better classification of the patients who died postoperatively in-hospital.

Weight-elimination had only a small impact on the classification performance for the combination of training using an artificial training set with a mortality rate of 30 percent, followed by testing the ANN performance on artificial test sets with 20 percent mortality rates (tr30/te20). Although the mean sensitivity rate of the weight-elimination network is higher than that of the ANN without weight-elimination, the other performance measures remained approximately the same, regardless of the cost function used.

In summary, there were inconsistent improvements for the mean specificity, the PPV, the area under the ROC curve, and the CCR. Despite these results, the mean sensitivity and the mean PNV were consistently the same or better when the ANN employed the weight-elimination cost function as opposed to simply using the standard backpropagation algorithm with the sum of squared errors cost function.

5.6 Evaluation of Variable Selection Using Weight-Elimination

Recall that the purpose of the weight-elimination cost function is to prune the input variable set to only those variables with the greatest impact on the desired outcome. It shrinks the weights of the least influential variables down to zero, thereby leaving the most important input variables to be easily discernable from the rest. This is, of course, an iterative process. With each pass of the data through the network, certain relationships are strengthened, while others are reduced.

The extraction of the weights can be most easily achieved using a single-layered network because there is only one weight per input variable. This process can also be achieved with double-layered networks, however, it is slightly more complex. In order for a variable to be considered unimportant in an ANN with a hidden layer, the weights of all of the nodes attached to that variable must be reduced to zero or near zero. In other words, if there are eight nodes in the hidden layer, and seven of the eight weights from the input variable are zero, this variable cannot be eliminated because one node still uses information from the variable to aid in the output classification.

Of course, the value of the weight-elimination scale constant, w_0, which determines how small a weight must be to be considered small, also influences whether a variable will be eliminated. If w_0 is too large, it will have no effect on the network, and if it is too small, all of the weights will be reduced to zero.

An examination of the weights in the hidden layer of the best-performing networks (found in Appendix H) showed that, in fact, no variables were eliminated. There were several cases where the weights of all but one node were reduced to zero, however, none were clearly removed from consideration.

At first, this result was surprising, however, after some thought, there were some probable explanations. First of all, the weight-elimination cost function did have a visible effect on the network's performance. As shown in Section 5.5 (c), weight-elimination increased the mean sensitivity of the ANN in all cases except when the network was trained on a lower prevalence of mortality than the test sets. This proves that weight-elimination is an effective approach to improving the correct classification rate of the nonsurvivors of CABG surgery. Second, the initial input variable list was based on Pliam *et al.*'s univariate analysis that selected the most influential variables from the list of potential risk factors. Knowing this, the input list had already been "preprocessed" to include the most important variables from the beginning. Despite an attempt to include additional risk factors that other research groups had found to influence the outcome, due to either a great majority of missing values or the complete absence of such a variable, it was not easy to find more factors that better identify patients at a higher risk of death.

There was a bias introduced by only using the variables found to be significant according to Pliam *et al.*'s work. It appears that the statistical analysis performed by the SFHI research group did in fact identify influential risk factors. Put another way, the weight-elimination technique agrees with the most important variables identified using univariate analysis. This is an interesting finding, since it supports the variable selection approach of weight-elimination. To verify this finding, however, the two approaches should be compared using a set of variables where there are some obvious variables with low correlation to the output that could be eliminated by each technique.

Another consideration is the fact that the ANNs appeared to "get stuck" after a certain number of iterations, and then were no longer able to learn more about the patterns. From the CCR and ASE curves in Appendix E, it is clear that at a certain point, the ANN stopped learning. This also meant that the weight-elimination function could no longer influence the network. Perhaps if the network had been able to learn more from the information in the CABG database, weight-elimination could have further reduced the weights of the less important variables. It is also possible that the weight-elimination scale

constant was not small enough to have an impact on the weights. However, a wide range of scale constants were tested and these values achieved the best performance.

5.7 Comparison with Previous Work

This section is important to validate the results found using the weight-elimination ANN to predict in-hospital mortality for CABG surgery patients. By comparing the results against Trigg's work, it may be possible to deduce the generalizability of this process to other medical databases. Then, a comparison against the results found by the SFHI research group will help to evaluate the effectiveness of this ANN model with respect to current mortality risk analysis techniques.

5.7 (a) Comparison with Trigg's Experiments

To summarize Trigg's [1997] work once more, she used an adult ICU database with 883 postoperative patients from the DECH in Fredericton, NB, to predict whether a patient would require mechanical ventilation for more than eight hours. The ICU database contained continuous and binary input variables with a binary output variable. The *a priori* distributions of the database were approximately 30-70 for patients requiring more than eight hours of artificial ventilation versus those that did not. Trigg [1997] used CCR as her measure of optimal performance, but also evaluated the techniques using ROC curves, number of iterations required to achieve maximum CCR, a constant predictor, the relative transinformation content (an information theory approach), and a minimum-distance classifier. The experiments that are relevant for comparison with the work performed in this thesis are those comparing single- and double-layered ANNs with and without weight-elimination with the regular data presentation technique.

Trigg [1997] found that weight-elimination improved the CCR (and overall performance) compared to their no weight-elimination counterparts. Double-layered ANNs slightly outperformed the single-layered networks, although the increased complexity meant a longer training time to achieve optimal performance. Trigg [1997] used the gearshifting technique suggested by Weigend *et al.* [1990a, 1990b] which reduced the time required to reach the maximum CCR. These experiments also showed that the weight-elimination networks actually took longer to reach the optimal performance than the ANNs without weight-elimination. Overall, the weight-elimination networks with one hidden layer

outperformed all other comparison techniques. Table 5.7 shows Trigg's ICU results for this network for comparison against the CABG simulations.

The experiments carried out in this thesis were on the CABG database from the SFHI in Daly City, CA, to predict in-hospital mortality. As stated previously, the true mortality rate is 3.7 percent — a highly skewed distribution (and the artificial test sets had a mortality rate of 20 percent). All of the input variables were binary except age, date of surgery and ejection fraction. Since nonsurvivors are the most difficult to classify, the focus of the performance measures was on the sensitivity rate, since this considers the correct classification of nonsurvivors. Here, the results showed that weight-elimination networks had a higher sensitivity rate than ANNs without weight-elimination. Using the CABG database, double-layered ANNs showed significant improvements in the sensitivity rate over single-layered networks that in some cases could not even learn any of the patterns. The gearshifting technique applied to the SFHI database offered little advantage over simply initializing the weight-elimination constant at a non-zero value before training began. In fact, even when gear-shifting was used, it was activated after only 10 epochs. Table 5.7 compares Trigg's best-performing network with the two best-performing networks from the experiments with the CABG database, tr20/teorig and tr30/te20.

Table 5.7: Comparison of double-layered ANN performance with weight-elimination on the adult ICU and CABG databases[8]

Performance Measure	Adult ICU database [Trigg 1997]	CABG database (tr20/teorig)	CABG database (tr30/te20)
CCR (%)	91.8	91.94 ± 0.79	76.74 ± 0.74
Sensitivity (%)	84.7 (72/85)	43.55 ± 7.59	61.68 ± 11.85
Specificity (%)	94.7 (198/209)	93.83 ± 0.77	79.82 ± 0.82
C-index	0.9301 ± 0.0195	0.8978 ± 0.0070	0.7535 ± 0.0057
Constant predictor (%)	71.1	96.2	80

As shown in Table 5.7, although the CABG database experiment with tr20/teorig has approximately the same CCR and specificity as the adult ICU simulation, its mean sensitivity is significantly lower (nearly half the sensitivity rate for the ICU database). Again, the other aspects of the adult ICU and tr20/teorig simulations are quite similar. Comparisons between the experiments with the artificial CABG datasets (tr30/te20) and the ICU database show a lower level of performance for all measuring techniques. As stated previously, the lower performance levels for the experiments with

[8] The results presented here for the CABG database are the means and standard deviations over 31 test sets derived from an original test set. All test sets within each group (teorig and te20) are the same size with the same prevalence. The results from these individual test sets are available in Appendix G.

training using the higher prevalence training sets are not necessarily indicative of their clinical relevance, since they achieve better classification of the nonsurvivors.

Trigg [1997] also managed to extract the input variables that had the greatest effect on the classification of patients as requiring more than eight hours of mechanical ventilation or not. She identified variables that influenced the decision to categorize the patients into the two output classes using weight-elimination on a single-layered ANN with the high-low node presentation approach. As stated in Section 5.6, this technique was unsuccessful with the CABG database.

One of the main objectives of this thesis was to use the weight-elimination cost function to reduce the number of input variables to only those that were most influential. Unfortunately, no variables were eliminated in any of the CABG experiments. Although it appears that weight-elimination was more effective on the ICU database, keep in mind that the ICU and CABG databases are looking at two problems that have dramatically different knowledge contained in the inputs. Not only are they two different medical conditions, they also differ in the type of data presented to the ANN (continuous and binary versus binary only data). It is important to emphasize the differences between the ICU and CABG databases and attempt to pinpoint reasons for the variations in performance.

First, successful implementation of a technique for one database does not mean that it is equally well applicable to other databases. One technique may be more suitable for a particular situation given the type of data available (i.e., continuous or categorical data, medical or financial data, etc.), or the best fitting function for the output (i.e., linear or nonlinear function). Richard & Lippmann [1991] and Lippmann & Shahian [1997] found equivalences or near-equivalences in the performance of ANNs compared with logistic regression for CABG databases. Richard & Lippmann [1991] used backpropagation ANNs with the sum of squared errors cost function, and Lippmann & Shahian [1997] trained their networks with stochastic gradient descent with early stopping, as opposed to the backpropagation algorithm with the weight-elimination cost function that was used in this thesis.

Second, intrinsic factors like the impact of the selected risk factors play a major role in the ANN's ability to develop an appropriate model. It might be that the DECH ICU database contains certain variables that have a great decisive impact on the model's output. For example, if a factor is present for a patient, that patient can be categorized into a particular output class with great certainty. As mentioned previously, CABG databases generally lack this type of information that "clearly" identifies a patient as

a survivor or a nonsurvivor [Orr 1997]. It is possible that in a CABG database, there are two patients with the same profile, yet one patient survives the surgery and the other does not. In this case, there is no clear evidence as to why the second patient did not survive the procedure.

Taking a closer look at the lambda values used in the experiments found in Table 5.4, one notices that the lambda values are quite close to zero (0.0001 to 0.0004). This indicates that the weight-elimination term had only a small part to play in the error cost function. Even Trigg [1997] had only marginal CCR improvements of one or two percent for networks using weight-elimination as opposed to the standard backpropagation algorithm. Also, the values of lambda used by Trigg [1997] were only 0.0003 or 0.0005, which are also quite close to zero, meaning that the influence of weight-elimination was minimal. Clearly, weight-elimination affected the classification ability of the networks applied to the CABG database as outlined in Section 5.5 (c).

Although weight-elimination did not eliminate any variables for the CABG database, one should not conclude that the weight-elimination technique is invalid. In this case, its ability to improve the mean sensitivity rate was demonstrated with the CABG experiments, and the CCR improvement for the ICU database allows the conclusion that weight-elimination is indeed a valid approach. In fact, its ability to eliminate unnecessary variables could be useful in a situation where the risk factors are unknown. In each of the above situations, clinicians already had a relatively good idea of what the risk factors for the particular outcome were, and hence only those variables were present in the data used in these experiments. In the case of the CABG surgery database, the input variables were chosen to be those selected by univariate analysis by the SFHI research group. It is possible that the "preprocessing" may have eliminated variables that weight-elimination may have found important and that statistical analysis did not recognize.

5.7 (b) Comparison with Pliam *et al.*'s Experiments

Although the current work was performed on the SFHI's cardiac database, the experiments from this thesis used only a subset of the data used by the SFHI research group. The SFHI research group used patient data spanning all of the years of data collection. Due to the technical advances that occurred in 1990 and 1991 [Shaw 1999], however, this is not the most homogenous dataset. Therefore, the simulations carried out in this thesis used the data from 1986 to 1991.

The majority of the mortality risk factors used in the current work were identified by Pliam *et al.* [1997] using univariate analysis. As noted previously, it is possible that using the influential variables chosen by Pliam and his colleagues as the basis for the initial variable list may have excluded variables that only the ANN could identify as relevant to the outcome. The only variables that the SFHI group had found to be significant that were not used in the thesis simulations were those related to smoking history. The reason for this was that there was a large number of missing values for this variable. Pliam and his colleagues had dealt with this problem by using three binary variables: current smoker, previous smoker and nonsmoker. There still arose many situations where all three of these factors were absent. In an effort to only use the most reliable data possible, these variables were omitted from consideration in the ANN experiments. Given that smoking history is a known risk factor for cardiac disease and that smoking (current or previous) can degrade a person's health, this is potentially an important risk factor, and hospitals should take care to consistently collect this information.

Of the variables used in this particular analysis, there were four input variables that Pliam *et al.* [1997] had not considered: date of surgery, ventricular aneurysm, mitral valve regurgitation, and unstable angina. These factors had been considered in several previous risk models, so the objective was to observe their impact on the ANN performance. It is common practice to include variables in an analysis that are not necessarily statistically significant in the current database, but have been consistently identified by other researchers to have a clinically known impact on the outcome [Higgins *et al.* 1992].

Table 5.8 contrasts the results achieved in this thesis work using ANNs with weight-elimination (trorig/teorig, tr20/teorig, tr10/te20, tr20/te20, and tr30/te20) with those found by Pliam and his colleagues using statistical risk model approaches (Parsonnet, Cleveland Clinic, Bayesian, and logistic regression). The mean predicted mortality rates of the ANNs are influenced by the *a priori* statistics of the training set. When the ANN was trained on a higher prevalence, the predicted mortality rate of the test sets was overestimated. Training on a lower mortality rate caused an underestimation of the test set mortality rate. Training on a dataset with a different prevalence than the test sets will always affect the way the network classifies the test sets. The Parsonnet model was developed on a training set with a higher mortality rate (8.9 percent [Parsonnet *et al.* 1989]) than the mortality rate of the test set (4.0 percent [Pliam *et al.* 1997]), so this information explains why the Parsonnet model has the highest mean predicted mortality for the statistical models.

Table 5.8: Comparison of risk models from current work and by SFHI research group

Researchers	Risk model	C-index[9]	Mean predicted mortality (%)	Actual test set mortality rate (%)
ANNs with weight-elimination	trorig/teorig	0.9387 ± 0.0029	2.04 ± 0.40	3.8
	tr20/teorig	0.8978 ± 0.0070	8.64 ± 0.71	3.8
	tr10/te20	0.7809 ± 0.0042	11.07 ± 0.46	20
	tr20/te20	0.7285 ± 0.0052	23.58 ± 0.54	20
	tr30/te20	0.7535 ± 0.0057	25.76 ± 0.72	20
Pliam *et al.* [1997]	Parsonnet	0.80 ± 0.02	9.0 ± 8.0	4
	Cleveland Clinic	0.80 ± 0.02	6.0 ± 6.0	4
	SFHI Bayesian	0.83 ± 0.02	7.6 ± 15.6	4
	SFHI logistic regression	0.80 ± 0.02	5.1 ± 7.7	4

Richard & Lippmann [1991] suggest a technique to eliminate the impact of the differences in mortality rates between the training and test sets, as outlined in Section 2.5. The approach involves multiplying the raw outcome data from the test sets by the true mortality rate of the dataset and dividing by the training set mortality rate. For tr20/teorig, this would mean multiplying the test data by 3.8 and dividing by 20. This calculation would give a mean predicted mortality rate of 1.7 percent compared with the true mortality rate of 3.8 percent. Although this figure is closer to the true mortality rate, it causes an underestimation of the test set mortality that is worse than when training on the true prevalence. The same thing happens for tr30/te20 (25.4 x 20 ÷ 30 = 16.9 percent). For tr10/te20, the mean mortality rate is slightly overestimated (10.6 x 20 ÷ 10 = 21.2 percent). Overall, this may not be a good "normalization" technique for this particular situation, because it reverses the desired effect of classifying more patients as nonsurvivors (with the intent that more of the nonsurvivors will be correctly classified).

From Table 5.8, the SFHI logistic regression model appears to be the best since its mean predicted mortality rate is the closest to the actual mortality rate for that test set (a 1.1 percent difference). The trorig/teorig ANN model has the next closest mean predicted mortality rate compared with its true rate (a difference of 1.8 percent). Of Pliam *et al.*'s risk models, the Bayesian model based on the SFHI data has the greatest area under the ROC curve indicating that it can better distinguish between survivors and nonsurvivors of CABG surgery than the other statistical models. Although the ANNs tested on the true mortality rate have greater areas under the ROC curves than all of the other models (statistical and ANNs), keeping in mind the emphasis on sensitivity for the ANN models is important. The ANNs would have had smaller C-indices if the measurements had been taken at the operating point for the sensitivity and specificity curves of the models. The focus of this thesis was the

[9] The ANN results presented here are the means and standard deviations over 31 test sets derived from an original test set. All test sets within each group (teorig and te20) are the same size with the same prevalence. The results from these individual test sets are available in Appendix G.

correct classification of the patients who do not survive CABG surgery while minimizing the misclassification of survivors. Therefore; finding a balance between sensitivity, specificity and CCR was the objective for these experiments.

Given the results presented in Table 5.8, determining the best mortality risk model is difficult since both sets of experiments focus on different outcomes (statistical models — model operating point, ANN models — highest sensitivity). Despite this, the ANN models with weight-elimination were within the same deviation from the true mortality rates as the statistical models, and the ANNs tested on the actual mortality rate had the greatest areas under the ROC curves.

5.8 General Remarks Regarding CABG Databases

In anticipation of improving the ANN modelling process, artificial training sets with different mortality distributions were employed. This approach proved to provide higher mean sensitivity rates, however, with minor reductions in the other performance measures. As mentioned earlier, increasing the sensitivity rate means that more of the difficult cases (the nonsurvivors) are correctly classified.

One problem involved in classifying CABG patients, with the data skewed towards the survivors, was the desire of the ANN to classify everything like a constant predictor does – according to the *a priori* statistics. Of course, this action would always result in a specificity of 100 percent, and a sensitivity of zero. Obviously, a model with such a performance level is useless for clinical decision making.

Trigg [1997] alluded to the problem of poor classification results when she applied the network to the classification problem of identifying postoperative ICU patients, who required more than 24 hours of artificial ventilation. The *a priori* probability for this outcome was 13 to 15 percent, and the ANN correct classification performance was actually slightly worse than the constant predictor (CCRs of 87.1 versus 87.4 percent, respectively). Trigg [1997] questioned whether it was a lack of a sufficient number of sample cases that prevented the network from surpassing the performance of the constant predictor, since her ICU database had only 883 patients (589 training patterns). As a final thought, Trigg [1997] also considered the possibility that the performance was detrimentally affected by the *a priori* statistics that were highly skewed towards 100 percent [Trigg 1997].

The original hypothesis for the study of the SFHI coronary artery surgery database was that a higher-than-normal prevalence for the training set would translate into better performance for test sets with lower mortality rates. The result was a higher number of nonsurvivors, who were correctly classified. However, these outcomes were paired with a higher misclassification rate of the survivors, and ultimately a lower overall CCR of the model.

In summary, coronary artery surgery is a difficult domain for mortality risk prediction. This fact is reinforced by the challenges that all medical researchers in this area have had, regardless of the size of their database. It is possible that "strong" predictors have yet to be found for CABG surgery patients. This may mean an expansion of the data collection variables to include factors that may have an impact on the outcome. An interaction between variables can make some of the predictions quite surprising and unexpected [Baxt 1994a]. Now is the time for researchers to turn their focus from the application of ANNs to various medical settings to the development of performance improvement techniques. It is no longer enough to simply state that applications of ANNs to the data that are highly skewed towards one output cannot achieve good correct classification rates. A greater effort should be made to break out of the mould, and look for means to overcome the classification problem of having an underrepresented output class.

Chapter 6: Concluding Remarks

This final chapter presents a summary of the findings from the experimental simulations using a feedforward backpropagation-trained ANN with the weight-elimination cost function applied to the problem of mortality prediction for patients undergoing CABG surgery at the SFHI in Daly City, CA. Once the contributions to new knowledge realized by this investigation are clearly highlighted, the thesis concludes with suggestions for related aspects of this work that should be further investigated to explain some relevant unanswered questions.

6.1 Conclusions

The initial objective of this thesis was to determine how changing the distribution of the training and test sets, by increasing the representation of the underrepresented class, would affect the ANN's classification performance. Training the ANN with a higher prevalence than in the test sets (30 percent mortality versus 20 percent mortality in the test sets) produced higher sensitivity values (correct classification of those who died). On the other hand, lower sensitivity levels were achieved using a training set with a lower prevalence than the artificial test sets (10 percent mortality versus 20 percent mortality in the test sets). This finding agrees with Baxt & Skora [1996] where they also achieved better correct classification when the network was trained on a higher-than-normal prevalence than found in the test sets. It was possible to make the same conclusion when testing the network on artificial test sets with a 20 percent mortality rate. These results highlight the significance of incorporating the *a priori* probabilities of the training set with the model's performance.

The second objective of the thesis was the extraction of the most important risk factors to predict survival. Examination of the resultant weights showed, however, that no variables were actually eliminated from the network. Obviously, the networks were affected by the weight-elimination cost

function, because of the difference in performance between the weight-elimination and no weight-elimination networks under the same conditions. Unfortunately, this aspect of the thesis objectives could not be achieved with the current database. It had been hypothesized that the same input variables would be selected as the least important for all of the training set/test set combinations under consideration.

6.2 Contributions to New Knowledge

The following is a summary of the contributions to new knowledge that resulted from the completion of this thesis.

1. Showed that the application of a feedforward backpropagation ANN to the estimation of in-hospital mortality following CABG surgery is a valid approach. This objective was clearly achieved given the results presented in Chapter 5 where the ANN's ability to correctly classify patients who died was exhibited in the sensitivity rates of each network under investigation.
2. This investigation was the first application of a feedforward backpropagation ANN using the weight-elimination cost function to in-hospital mortality prediction for a CABG patient database.
3. Demonstrated that the weight-elimination cost function added to the SSE error term improves the ANN's ability to classify nonsurvivors. Improved performance was defined as having a higher sensitivity rate because the nonsurvivors are more difficult to classify. Weight-elimination improved the ANN's classification performance. From the ANN analysis, the model's sensitivity is the same or improves when using the weight-elimination technique.
4. Demonstrated how training with an artificial dataset with a higher-than-normal prevalence improves the test set classification rate of nonsurvivors (i.e., achieved a higher sensitivity rate).
5. Showed that training on a dataset with a lower prevalence than the test sets causes worse nonsurvivor classification performance (i.e., resulted in a lower sensitivity rate).
6. Discovered that although the weight-elimination cost function did not entirely eliminate any input variables from consideration, the weight-elimination technique did achieve a noticeable improvement over ANNs trained without weight-elimination.
7. Demonstrated that double-layered ANNs outperform single-layered ANNs when estimating in-hospital mortality prediction of CABG patients by having a higher sensitivity rate, thereby correctly classifying more nonsurvivors.
8. Showed how "gear-shifting" the weight-elimination cost function to turn on after several hundred epochs did not always improve the sensitivity of the test sets.
9. This analysis of the SFHI CABG database generated the suggestion for hospitals to begin collecting information on the patients who are denied this type of cardiac surgery. Since patients who do not survive CABG surgery likely have many common characteristics as those patients who are denied surgery, the addition of this information to the CABG database could enrich the information regarding nonsurvivors, and possibly improve a risk model's performance.

6.3 Future Work

As a fitting conclusion to this research project, issues that arose as initial thesis objectives and which were eliminated to maintain a reasonable workload, during the literature survey, or while performing the experimental simulations, are listed as future work. These issues fall under four categories: ANN development, statistical technique developments, CABG database improvements, and new applications.

ANN Development:

- Develop a program that can automatically optimize all of the ANN parameters without requiring user supervision.
- Develop a guideline for optimizing the ANN architecture that is generalizable to various medical settings.
- Observe the sensitivity of the ANN to the removal of input variables whose weights were reduced to zero by the weight-elimination cost function for all but one of the weights.
- Apply the approach of modifying the distribution of the database to another medical database with a similar *a priori* distribution and observe the ANN's performance (i.e., attempt to predict mortality for the DECH ICU database).
- Investigate the possibility of using the optimal brain damage technique developed by Le Cun *et al.* [1990].

Statistical Technique Developments:

- Develop a new performance measure that is independent of the *a priori* statistics.
- Use statistical analysis to develop a logistic regression model using this reduced database that spans a shorter time period, and compare the results with those found in this thesis.
- Compare the variables that remain after employing statistical variable elimination techniques and an ANN with weight-elimination.
- Develop an ANN using variables deemed important by statistics, and compare the results with the most important variables identified using the weight-elimination cost function with ANNs.

CABG Database Improvements:

- Hospitals should start collecting data on patients who are denied surgery. Add this data to the CABG database as potential nonsurvivors to observe their impact on the model's performance.
- Medical researchers should continue to investigate other possible factors related to CABG mortality. The overall poor performance by all CABG risk models may indicate that the defining factors for CABG mortality still have not been identified.
- Medical researchers should try using input variables other than risk factors. Include pre-operative physiological variables, and where medication or medical equipment alter or normalize the variable, note the presence or absence of the medical intervention being used. Perhaps using continuous variables and those that are not risk factors may improve the ANN's classification performance.
- Apply the SFHI model to another CABG database, and compare the results.

- The SFHI database (and most other CABG databases) define mortality as death within a short period of time following surgery (i.e., in-hospital, within 30 days of surgery, within 30 days of discharge, etc.). There exists a possibility, however, that some of the patients who are identified as survivors may have actually died shortly after the "cutoff" point, but were still classified as survivors. Even if a patient dies several weeks or months after surgery, should that still be considered a "positive" outcome? By using only patients who survived, say one or more years following the surgical procedure, as survivors, perhaps the skewness of the data could be reduced. The rest of the patients would be considered nonsurvivors. Another approach to better identify the risk factors might be to use only patients who die within say 30 days of surgery as nonsurvivors, and then only patients who survive more than one year postsurgery as survivors. Perhaps the elimination of the patients who are "in-betweens" might better highlight the deciding factors between life and death.

New Applications:

- Replace the missing values for the smoking history variable (SMOKE2) by categorizing these as nonsmoking patients and observe the effect on the sensitivity of the model.
- Identify the appropriate risk factors to develop a model to predict the length of hospital stay following the surgery for the SFHI database.
- Use the SFHI cardiac database to develop a model to predict morbidity, since this is also an important consideration when evaluating the quality of care.

References

[Baxt 1991]
Baxt WG. Use of an artificial neural network for the diagnosis of myocardial infarction. Ann of Internal Medicine 1991;115:843-848.

[Baxt 1993]
Baxt WG. A neural network trained to identify the presence of myocardial infarction bases diagnostic decision on nonlinear relationships between input variables. Neural Comput Applic 1993;1:176-182.

[Baxt 1994a]
Baxt WG. A neural network trained to identify the presence of myocardial infarction bases some decisions on clinical associations that differ from accepted clinical teaching. Med Decis Making 1994;14:217-222.

[Baxt 1994b]
Baxt WG. Complexity, chaos and human physiology: the justification for non-linear neural computational analysis. Cancer Lett. 1994;77:85-93.

[Baxt & Skora 1996]
Baxt WG, Skora J. Prospective validation of artificial neural network trained to identify acute myocardial infarction. Lancet 1996;347:12-15.

[Blum 1992]
Blum A. Neural Networks in C++. John Wiley & Sons, Inc.: New York, 1992.

[Buchman *et al.* 1994]
Buchman TG, Kubos KL, Seidler AJ, Siegforth MJ. A comparison of statistical and connectionist models for the prediction of chronicity in a surgical intensive care unit. Crit Care Med 1994;22(5):750-762.

[Burke *et al.* 1994]
Burke HB, Rosen DB, Goodman PH. Comparing artificial neural networks to other statistical methods for medical outcome prediction. Proc of International Conf on Neural Networks 1994 (ICNN'94), IEEE. Orlando, FL, USA: June 26-July 2, 1994.

[Buskard 1994]
Buskard TT. Estimating duration of ventilation, length of stay, and mortality in the intensive care unit using artificial neural networks. Master's thesis, Dept of Electrical Engineering, University of New Brunswick 1994.

[Buskard *et al.* 1994]
Buskard T, Stevenson M, Frize M, Solven F. Estimation of ventilation, length of stay, and mortality using artificial neural networks. Proceedings of the 1994 Canadian Conf on Electrical & Computer Engineering (CCECE'94) September 1994:726-729.

[Clark *et al.* 1994]
Clark RE, Edwards FH, Schwartz M. Profile of preoperative characteristics of patients having CABG over the past decade. Ann Thorac Surg 1994;58:1863-1865.

[Clark 1996]
Clark RE. Calculating risk and outcome: The Society of Thoracic Surgeons database. Ann Thorac surg 1996;62:S2-S5.

[CVOR 1997]
Consortium for Virtual Operations Research (CVOR). Boutell T, maintainer. West Virginia University, Penn State University:http://cvor.pe.wvu.edu/faq/nnfaq.htm. 1997.

[Ebert 1989]
Ebert PA. Keynote address from the American College of Surgeons, Chicago, IL. Circulation 1989;79(Suppl 1):I2.

[Edwards *et al.* 1987]
Edwards FH, Graeber GM. The theorem of Bayes as a clinical research tool. Surgery, Gynecology & Obstetrics 1987;165:127-129.

[Edwards *et al.* 1988]
Edwards FH, Albus RA, Zaitchuk R, et al. Use of a Bayesian statistical model for risk assessment in coronary artery surgery. Ann Thorac Surg 1988;45:437-440.

[Edwards *et al.* 1994a]
Edwards FH, Clark RE, Schwartz M. Coronary artery bypass grafting: The Society of Thoracic Surgeons national database experience. Ann Thorac Surg 1994;57:12-19.

[Edwards *et al.* 1994b]
Edwards FH, Clark RE, Schwartz M. Practical considerations in the management of large multiinstitutional databases. Ann Thorac Surg 1994;58:1841-1844.

[Edwards *et al.* 1997]
Edwards FH, Grover FL, Shroyer ALW, Schwartz M, Bero J. The Society of Thoracic Surgeons national cardiac surgery database: current risk assessment. Ann Thorac Surg 1997;63:903-908.

[Ennett & Frize 1998]
Ennett CM, Frize M. An investigation into the strengths and limitations of artificial neural networks: an application to an adult ICU patient database. Proc of AMIA'98 Fall Symposium, Orlando, FL, Nov 7-11, 1998:998.

[Ennett *et al.* 1999]
Ennett CM, Frize M, Shaw RE. Methodologies for predicting coronary surgery outcomes. Proc First Joint Meeting BMES-EMBS, Atlanta, GA, Oct 13-16, 1999.

[Fausett 1994]
Fausett L. Fundamentals of Neural Networks. Prentice-Hall: Englewood Cliffs (NJ), 1994.

[Forsström & Dalton 1995]
Forsstrom JJ, Dalton KJ. Artificial neural networks for decision support in clinical medicine. Annals of Medicine 1995;27:509-517.

[Frize *et al.* 1995]
Frize M, Solven FG, Stevenson M, Nickerson BG, Buskard T, Taylor K. Computer-assisted decision-support systems for patient management in an intensive care unit. Proc of Medinfo'95 1995:1009-1012.

[Frize *et al.* 1996]
Frize M, Solven FG, Stevenson M, Nickerson BG, McGowan HCE. Information technologies approach and development for various medical applications. Proc of the 1996 Canadian Conference on Electrical & Computer Engineering (CCECE'96) 1996: 365-368.

[Frize *et al.* 1997]
Frize M, Trigg HCE, Solven FG, Stevenson M, Nickerson BG. Decision-support systems designed for critical care. Proceedings of the 1997 AMIA Fall Symposium. 1997:855.

[Frize *et al.* 1998]
Frize M, Wang L, Ennett CM, Nickerson B, Solven FG, Stevenson MH. New advances and validation of knowledge management tools for critical care using classifier techniques. Proceedings of the 1998 AMIA Fall Symposium. 1998: 553-558.

[Grover *et al.* 1995]
Grover FL, Hammermeister KE, Shroyer ALW. Quality initiatives and the power of the database: what they are and how they run. Ann Thorac Surg 1995;60:1514-1521.

[Grover *et al.* 1996]
Grover FL, Shroyer ALW, Hammermeister KE. Calculating risk and outcome: The Veterans Affairs Database. Ann Thorac Surg 1996;62:S6-S11.

[Higgins *et al.* 1992]
Higgins TI, 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 1992;267:2344-2348.

[Hornik *et al.* 1989]
Hornik K, Stinchcombe M, White H. Multilayer feedforward networks are universal approximators. Neural Networks 1989;2:359-366.
[Itchhaporia *et al.* 1996]
Itchhaporia D, Snow PB, Almassy RJ, Oetgen WJ. Artificial neural networks: Current status in cardiovascular medicine. J Am Coll Cardiol 1996;28(2):515-521.
[Kattan & Beck 1995]
Kattan MW, Beck JR. Artificial neural networks for medical classification decisions. Arch Pathol Lab Med 1995;119:672-677.
[Keon & Menzies 1992]
Keon WJ and Menzies SC. Morbidity and mortality after myocardial revascularization in patients with ischemic heart disease, In: Quality of Life After Open Heart Surgery. Ed: Walter PJ. Kluwer Academic Publishers: Dordrecht. 1992:107-114.
[Kleinbaum *et al.* 1998]
Kleinbaum DG, Kupper LL, Muller KE, Nizam A. Applied Regression Analysis and Other Multivariable Methods. Duxbury Press: Pacific Grove. 1998:pp.656.
[Krogh & Hertz 1992]
Krogh A, Hertz JA. A simple weight decay can improve generalization. In: Lippmann RP, Moody J, Touretzky, eds. Advances in Neural Information Processing Systems 4 (NIPS'91). San Matteo, CA: Morgan Kaufmann, 1992:950-957.
[Le Cun *et al.* 1990]
Le Cun Y, Denker JS, Solla SA. Optimal brain damage. In: Touretzky D, ed. Advances in Neural Information Processing Systems 2 (NIPS'89), Denver, CO: Morgan Kaufman, 1990:598-605.
[Lippmann *et al.* 1995]
Lippmann RP, Kukolich L, Shahian D. Predicting the risk of complications in coronary artery bypass operations using neural networks. In: Tesaukro G, Touretzky D, Leen T, eds. Advances in Neural Information Processing Systems 7 (NIPS'94). San Matteo, CA: Morgan Kaufmann, 1995:1055-1063.
[Lippmann & Shahian 1997]
Lippmann RP, Shahian DM. Coronary artery bypass risk prediction using neural networks. Ann Thorac Surg 1997;63:1635-1643.
[Livingstone *et al.* 1997]
Livingstone DJ, Manallack DT, Tetko IV. Data modelling with neural networks: Advantages and limitations. J Computer-Aided Molecular Design, 1997;11:135-142.
[Masters 1993]
Masters T. Practical Neural Network Recipes in C++. Academic Press: San Diego, 1993.
[McGowan *et al.* 1996]
McGowan HCE, Stevenson M, Frize M, Solven FG. A reporting guideline for medical applications of artificial neural networks. Proc of CMBEC 22, June 1996:16-17.
http://www.sce.carleton.ca/faculty/frize/charpap.htm
[Miller 1977]
Miller DW, Jr. The Practice of Coronary Artery Bypass Surgery. Plenum Medical Book Company: New York. 1977.
[Nugent 1995]
Nugent WC. Discussion on article by Orr *et al.* [1995b]. Arch Surg 1995;130:306.
[O'Connor *et al.* 1992]
O'Connor GT, Plume SK, Olmstead EM, Coffin LH, Morton JR, Maloney CT, Nowicki ER, Levy DG, Tryzelaar JF, Hernandez F, *et al.* Multivariate prediction of in-hospital mortality associated with coronary artery bypass graft surgery. Northern New England Cardiovascular Disease Study Group. Circulation 1992 Jun;85(6):2110-2118.
[Ohno-Machado *et al.* 1998]
Ohno-Machado L, Fraser HS, Ohrn A. Improving machine learning performance by removing redundant cases in medical data sets. Proc of AMIA'98 Fall Symposium. 1998:523-527.

[Orr *et al.* 1995a]
Orr RK, Kantor WN, Maini BS, Sottile F. Using a neural network to predict cardiac surgery ICU length of stay (abstr). Crit Care Med 1995;23:A137.
[Orr *et al.* 1995b]
Orr RK, Maini BS, Sottile FD, Dumas EM, O'Mara P. A comparison of four severity adjusted models to predict mortality after coronary artery bypass graft surgery. Arch Surg 1995;130:301-306.
[Orr 1997]
Orr RK. Use of a probabilistic neural network to estimate the risk of mortality after cardiac surgery. Med Decis Making 1997;17(2):178-185.
[Parsonnet *et al.* 1989]
Parsonnet V, Dean D, Bernstein AD. A method of uniform stratification of risk for evaluating the results of surgery in acquired adult heart disease. Circulation 1989;79(Suppl 1):I3-I12.
[Penny & Frost 1996]
Penny W, Frost D. Neural networks in clinical medicine. Med Decis Making 1996;16:386-398.
[Pliam *et al.* 1997]
Pliam MB, Shaw RE, Zapolanski A. Comparative analysis of coronary surgery risk stratification models, J Invas Cardiol 1997;9:203-222.
[Richard & Lippmann 1991]
Richard MD, Lippmann RP. Neural network classifiers estimate Bayesian *a posteriori* probabilities. Neural Computation 1991;3(4):461-483.
[Rumelhart *et al.* 1986a]
Rumelhart DE, Hinton GE, Williams RJ. Learning internal representations by error propagation. In DE Rumelhart, JL McClelland, eds, Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol 1:318-362. MIT Press: Cambridge, MA, 1986.
[Rumelhart *et al.* 1986b]
Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature 323:533-536.
[Shaw 1999]
Shaw RE. Personal communication dated Aug 25, 1999.
[Shroyer *et al.* 1998]
Shroyer ALW, Grover FL, Edwards FH. 1995 coronary artery bypass risk model: the Society of Thoracic Surgeons adult cardiac national database. Ann Thorac Surg 1998;65:879-884.
[Shroyer *et al.* 1999]
Shroyer ALW, Plomondon ME, Grover FL, Edwards FH. The 1996 coronary artery bypass risk model: The Society of Thoracic Surgeons adult cardiac national database. Ann Thorac Surg 1999;67:1205-1208.
[Smith 1993]
Smith M. Neural Networks for Statistical Modeling. Van Nostrand Reinhold: New York, 1993.
[Specht 1990]
Specht DF. Probabilistic neural networks. Neural Networks 1990;3:109-118.
[Tabachnick & Fidell 1989]
Tabachnick BG, Fidell LS. Using Multivariate Statistics, 2nd Edition. Harper & Row, New York:1989.
[Tong 1983]
Tong H. Threshold Models in Non-Linear Time Series Analysis. Springer, 1983.
[Trigg 1997]
Trigg HCE. An investigation of methods to enhance the performance of artificial neural networks used to estimate ICU outcomes. Master's thesis, Department of Electrical Engineering, University of New Brunswick: Fredericton, NB. 1997.
[Tu & Guerriere 1993]
Tu JV, Guerriere MRJ. Use of a neural network as a predictive instrument for length of stay in the intensive care unit following cardiac surgery. Comput Biomed Res 1993;26:220-229.
[Turner *et al.* 1995]
Turner JS, Morgan CJ, Thakrar B, Pepper JR. Difficulties in predicting outcome in cardiac surgery patients. Crit Care Med 1995;23:1843-1850.

[Vicino 1998]
Vicino F. Some reflections on artificial neural networks and statistics: two ways of obtaining solutions by working with data. Substance Use & Misuse, 1998;33(2):221-231.
[Warner 1997]
Warner BA. Thoughts and considerations on modeling coronary bypass surgery risk. Ann Thorac Surg 1997;63:1529-1530.
[Weigend *et al.* 1990a]
Weigend AS, Rumelhart DE, Huberman BA. Back-propagation, weight-elimination and time series prediction. In DS Tourestzky, JL Elman, TJ Sejnowski, GE Hinton, eds. Proc of the 1990 Connectionist Models Summer School. Morgan Kaufmann: San Mateo. 1990:105-116.
[Weigend *et al.* 1990b]
Weigend AS, Huberman BA, Rumelhart DE. Predicting the future: a connectionist approach. Int J of Neural Systems. 1990;1(3):193-209.
[Weigend *et al.* 1991a]
Weigend AS, Rumelhart DE, Huberman BA. Generalization by weight-elimination with application to forecasting. In RP Lippmann, J Moody, DS Touretzky, eds. Advances in Neural Information Processing Systems (NIPS*90), Volume 3. Morgan Kaufmann: San Mateo. 1991;3:875-882.
[Weigend *et al.* 1991b]
Weigend AS, Rumelhart DE, Huberman BA. Generalization by weight-elimination applied to currency exchange rate prediction. Proc of the Int Joint Conference on Neural Networks (IJCNN), Volume 1, Seattle, WA. 1991:837-841.
[Werbos 1974]
Werbos P. Beyond regression: new tools for prediction and analysis in the behavioral sciences. PhD Thesis, Harvard University, Cambridge, MA. August 1974.
[Woods & Bowyer 1997]
Woods K, Bowyer KW. Generating ROC curves for artificial neural networks. IEEE Trans on Medical Imaging 1997;16(3):329-337.
[Youngson 1992]
Youngson RM. Collins Dictionary of Medicine. Harper Collins Publishers: Glasgow. 1992: 36.

Appendices

Appendix A: Strengths and Limitations of Artificial Neural Networks

Originally published as "Investigation into the Strengths and Limitations of Artificial Neural Networks: An Application to an Adult ICU Patient Database" by Colleen M. Ennett and Monique Frize. Copyright ©1998 American Medical Informatics Association. Reprinted, with permission, from the Proceedings of the AMIA'98 Fall Symposium, Orlando, FL, Nov 7-11, 1998:998.

Investigation into the Strengths and Limitations of Artificial Neural Networks: An Application to an Adult ICU Patient Database

Colleen M. Ennett[1] and Monique Frize, PhD, PEng[1,2]
[1]University of Ottawa, [2]Carleton University, Ottawa, Ontario, Canada

The objective was to determine the optimal operating conditions for an artificial neural network (ANN) to estimate outcomes. The simulations involved using the 51 inputs while changing the desired output variable. Comparing the correct classification rate (CCR) of an ANN with that of a constant predictor (CP) results indicates the minimum number of sample patterns an ANN requires for minimally acceptable outcome estimation, and establishes the limitation of the ANN as a useful tool.

INTRODUCTION

Few medical researchers have achieved correct classification results with ANNs in the 90+% range [1]. This paper discusses how well a back propagation feed-forward ANN separates two output classes as the representation of the dominant class approaches 100%.

METHODOLOGY

We performed the simulations using the same database and neural network code as Trigg [1] with the data sorted into two categories: post-operative and nonpost-operative patients. We used the CCR and average squared error (ASE) to evaluate the performance of the network.

Six different situations involving the number of hours of mechanical ventilation were investigated (≤ 4 hrs, 12, 24, 36 and 336 hrs, and between 24 and 336 hrs), as well as estimates of the length of ICU stay (0, ≤ 1, 4, 5, and 14 days). A commonly estimated medical outcome is mortality (or "survival rate"), therefore, this output variable was also investigated. The objective was to observe the changes in CCR and ASE for each outcome.

RESULTS

From the simulation results, it appears that the ANN had more difficulty classifying the NONPOSTOP patients than the POSTOP patient cases. A possible explanation is the extreme diversity of the circumstances surrounding the patients in the NONPOSTOP subdatabase, making them more difficult to classify.

To see the relationship between the CCR of the CP and the ANN for the two databases, we plotted the results

of CCR versus the proportion of representative samples in the database. The results showed that the CCR of the ANN and the CP converge to a theoretical limit for the superior performance of the ANN. This occurs as the division between the two desired outputs becomes highly skewed towards 100% for one of the outcomes. In cases where the number of sample patterns for a particular case are quite small, after the first few experiments with the ANN, everything becomes classified as belonging to the largest class — in essence, the ANN becomes a CP.

Using linear regression of the CCR for the ANN, we approximately identified this limit. The point at which the linear regression line crosses the CP predictions would be the theoretical limit for the ANN. After this point, as the division between the output classes becomes more skewed, the ANN either becomes a CP or its CCR is lower than that of a CP. We discovered that the dominant output class may represent at most 85.9% of the POSTOP database, and 83.5% for the NONPOSTOP database under consideration.

CONCLUSION

For this adult ICU patient database, it seems that in order to have a notable improvement over a CP, the ANN requires the dominant output class to represent no more than 85.9% of the database for POSTOP patients, or 83.5% of the cases in a NONPOSTOP database. These limitations cannot necessarily be directly applied to other databases (medical or otherwise) because the ANN relies heavily on the relationship between the input parameters. However, this information could be used as a guideline for other applications.

References

1. Trigg HCE. An investigation of methods to enhance the performance of artificial neural networks used to estimate ICU outcomes. Master's Thesis, Department of Electrical Engineering, University of New Brunswick: Fredericton, NB: 1997.

Acknowledgement. The authors wish to extend sincere gratitude to Helena Ho, a fourth year engineering student at the University of Ottawa, for her substantial help in carrying out simulations.

Appendix B: Letters of Permission for Copyrighted Material

Permission for use of "Investigation into the strengths and limitations of artificial neural networks: an application to an adult ICU patient database" by CM Ennett and M Frize from the American Medical Informatics Association (AMIA).

From: "Ina King" <ina@mail.amia.org>
Organization: AMIA
To: Colleen Ennett <cennett@uottawa.ca>
Date: Mon, 2 Aug 1999 14:35:12 EST
Subject: Reprint or republication permission
X-Confirm-Reading-To: "Ina King" <ina@mail.amia.org>
X-pmrqc: 1
Priority: normal

Ms. Ennett,

I am the Administrative Coordinator at the American Medical Informatics Association (AMIA). I am also the executive assistant to Mr. Dennis Reynolds, Executive Director of AMIA.

Your request for permission to reprint or republish your paper, which is found in the 98 AMIA Proceedings was forward to my attention.

I have spoken with Mr. Reynolds. He grants permission for the reprint/republication, in conjunction with your thesis, of your paper entitled: "Investigation into the strengths and limitations of artificial neural networks: an application to an adult ICU patient database."

Please feel free to contact me if you need anything else.

Sincerely,

Ina King
Administrative Coordinator
American Medical Informatics Assoc.
Suite 401
4515 St. Elmo Avenue
Bethesda MD 20814
301-657-5916 direct
ina@mail.amia.org

Permission for use of "Methodologies for predicting coronary surgery outcomes" by CM Ennett, M Frize, and RE Shaw from the Institute of Electrical and Electronics Engineers (IEEE).

X-Sender: lnigro@pop.ieee.org
Date: Tue, 03 Aug 1999 09:24:04 -0400
To: cennett@uottawa.ca
From: Lisa Nigro <l.nigro@ieee.org>
Subject: * Copyright permission
Cc: b.hagen@ieee.org
X-MIME-Autoconverted: from quoted-printable to 8bit by cliff.Uottawa.Ca id JAA55880

3 August 1999

Ms. Colleen Ennett
2001-420 Gloucester Street,
Ottawa, ON, Canada

SUBJECT: "Methodologies for Predicting Coronary..." by Ennett et al.

Dear Ms. Ennett:

This is in response to your letter of 29 July in which you have requested permission to reprint, in your upcoming thesis, your above IEEE copyrighted paper. We are happy to grant this permission.

Our only requirement is that the following copyright/credit notice appears prominently on the first page of the reprinted paper, with the appropriate details filled in:

> © 19xx IEEE. Reprinted, with permission, from Proceedings of
> (full conference name; place and date of conf.; page numbers).

Sincerely yours,

William J. Hagen, Manager
IEEE Intellectual Property Rights
WJH:ln

Permission for use of "Use of an artificial neural network for the diagnosis of myocardial infarction" by William G. Baxt from the American College of Physicians.

A C P
A S I M

American College of Physicians

American Society of Internal Medicine

ACP–ASIM
190 N. Independence Mall West
Philadelphia, PA 19106-1572

215 351-2400
800 523-1546
www.acponline.org

August 10, 1999

WAA006070
Your reference #:Myocardial Infarction

Colleen Ennett
2001-420 Gloucester Street
Ottawa, Ontario K1R 7T7
CANADA

Dear Ms. Ennett:

Thank you for your request to reproduce the following from *Annals of Internal Medicine:*

> Appendix from Baxt, W. Use of an Artificial Neural Network for the Diagnosis of Myocardial Infarction. *Ann Intern Med.* 1991;115:843-848.

Permission is granted to reproduce the preceding material with the understanding that you will give appropriate credit to *Annals of Internal Medicine* as the original source of the material. Any translated version must carry a disclaimer stating that the American College of Physicians—American Society of Internal Medicine is not responsible for the accuracy of the translation. This permission grants non-exclusive, world-wide rights for this edition in printed format only. ACP—ASIM does not grant permission to reproduce entire articles or chapters on the Internet. This letter represents the agreement between ACP--ASIM and Colleen Ennett for request WAA006070 and supersedes all prior terms from the requestor.

Thank you for your interest in *Annals of Internal Medicine.* If you have any further questions or would like to discuss the matter further, please contact me at 215-351-2630.

Sincerely,

Joshua Roberts
Permissions Coordinator
ACP-ASIM

Appendix C: Baxt's Description of a Feedforward Backpropagation ANN

Segment taken from Appendix on page 847.

"The mathematical operation of the network ... is independent of its specific use and can be viewed as generic. The network functions by the application of binary or analog coded data comprising the pattern set to the ... input units. This signal is then multiplied by the initially random weights on the projections between each input unit and the first layer of hidden units:

$$net_{pi} = \sum_{j=0}^{n} w_{ij} a_{pj} + bias_i \qquad \text{C.1}$$

where net_{pi} equals the net input of the unit for pattern p, w is a random weight, a is the input value applied to the unit, j represents the input or presynaptic units, I represents the first layer hidden unit or postsynaptic unit, and bias is a modifiable weight that is multiplied by an input that is always equal to 1.

"The net activation of the hidden unit is calculated by:

$$a_{pi} = \frac{1}{1 + e^{-net_{pi}}} \qquad \text{C.2}$$

"The activation of the second layer hidden units is calculated by use of the first equation in a manner analogous to that used to calculate the activation of the first layer hidden units. In the latter instance, the input signal now becomes the net activation of each first layer hidden unit, this is multiplied by the weights on the projections between each of the first and second layer hidden units. Unit activation is calculated by the use of equation 2. Network output is also calculated in an analogous manner, with the second layer hidden units now providing the input signal, which is multiplied by the weights on the ten projections to the output unit. Unit activation (network output) is again calculated by use of equation 2.

"The difference between a training pattern output or target value and the network output a_{pi}, termed e, is calculated by subtracting network output from the target value t_{pi}. An e is calculated for each noninput unit of the network and used by the back propagation algorithm to modify all weights of the network such that, when pattern p is again inputted, the difference between network output and the pattern target value will diminish.

"Weight is modified by the derivation of delta. The delta, δ, for the output unit is calculated by:

$$\delta_{pi} = (t_{pi} - a_{pi}) f(net_{pi}) \qquad \text{C.3}$$

where $f_i(net_{pi})$ is the derivative of the activation function with respect to a change in the net input to the unit.

"The delta for the hidden units is calculated in terms of the units to which they project and the weights on those projections:

$$\delta_{pi} = f_i(net_{pi}) \sum_{k=0}^{n} \delta_{pk} w_{ki} \qquad \text{C.4}$$

"Weights and biases are updated by the calculation of the delta weight:

$$\Delta w_j = \alpha w_j + ((1 - \alpha)(\varepsilon \delta_i a_j)) \qquad \text{C.5}$$

where α is termed network momentum and ε is termed the learning rate parameter. Weights are updated by adding the delta weight to the old weight.

"Training of the network is followed by calculating the summed square of the error, represented by e, across the entire pattern set:

$$E = \sum_{p} \sum_{i} (t_{pi} - o_{pi})^2 \qquad \text{C.6}$$

where the index p ranges over the set of input patterns P, I ranges over the set of output units, t_{pi} is the target of pattern p, o_{pi} is the network output for pattern p. When the summed square of the error has ceased diminishing or has reached 0, the network has been trained. If no relationship between the pattern sets and their target value exists, this value will not diminish."

Appendix D: Methodologies for Predicting Coronary Surgery Outcomes

Originally published as "Methodologies for Predicting Coronary Surgery Outcomes" by Colleen M. Ennett, Monique Frize, and Richard E. Shaw.

Methodologies for Predicting Coronary Surgery Outcomes

C.M. Ennett†(cennett@uottawa.ca), M. Frize, PhD, PEng†‡, R.E. Shaw, PhD*
†School of Info Tech & Eng'g, 161 Louis Pasteur, University of Ottawa, Ottawa, ON, K1R 7T7
‡Systems & Computer Eng'g, 1125 Colonel By Dr., Carleton University, Ottawa, ON, K1S 5B6
*San Francisco Heart Institute, Seton Medical Center, 1900 Sullivan Ave., Daly City, CA, 94015

ABSTRACT
Preliminary results using an artificial neural network (ANN) on a coronary artery bypass grafting (CABG) surgery database highlighted challenges when faced with a low representation of a binary variable in the database. We will artificially alter the distribution of the database by reproducing or removing cases, and observe any changes in ANN performance. Final results will be presented at the conference.

INTRODUCTION
Useful mortality prediction models for patients undergoing CABG surgery are difficult to develop. To date, no one has designed a model with an area under the ROC curve of greater than 80 to 85 percent. This performance is somewhat lower than that of trained models in other areas of medical research such as acute myocardial infarction.

Pliam *et al.*'s research group from the San Francisco Heart Institute (SFHI) compared six different statistical and mathematical mortality risk models using their heart surgery patient database. They concluded these CABG models could accurately classify outcomes about 80 percent of the time [1].

METHODS
We applied a backpropagation feedforward ANN to the SFHI database using a technique called weight-elimination. This penalty function reduces the weights of the least important variables to zero, and selects the variables with the most significant impact on the output. We will compare our results with those of Pliam *et al.* [1].

RESULTS
Preliminary results identified several challenges of working with a CABG database. The most significant obstacle was the shortage of sample cases of nonsurvivors. Although a low mortality rate is desirable from a medical point of view, this small representative sample of nonsurvivors (less than 5 percent) poses a real problem when developing a prediction model for survival. With so few examples, ANNs have difficulty in recognizing the input patterns of nonsurvivors, thereby producing poorly generalizable results.

DISCUSSION
Although most patients requiring heart surgery receive it, those who are rejected may have much in common with the nonsurvivors. Data for unsuitable patients is not normally collected in a database. However, an unsuitable patient likely has very advanced co-morbid conditions that would make anesthesia dangerous, poor flow in the distal portion of the native vessel that cannot be rectified with endarterectomy, and/or very poor ejection fraction. Nonsurvivors of CABG surgery likely share some of these characteristics.

To increase the sample size of potential nonsurvivors, hospitals should begin collecting data about the patients who are unsuitable for CABG surgery. A larger dataset of these patients may improve the performance of mortality risk stratification models of all forms.

Since this information is not currently available, we will test two approaches to increase the number of cases of nonsurvivors artificially. One technique is to copy nonsurvivor cases until their representation in the database is 20 percent. The second is to remove survivor cases until the representation of nonsurvivors is 20 percent. In previous work, Ennett and Frize [2] showed that ANN performance on an intensive care unit (ICU) database diminished rapidly when the representation of a binary variable was less than 15 percent, so a minimum of 20 percent was chosen to include a factor of safety.

CONCLUSIONS
The challenge involved in the shortage of representative

samples for the nonsurvivors may be overcome by artificially changing the distribution of the database. This goal can be achieved by either reproducing the nonsurvivor cases until their representation is sufficiently large, or by removing survivor cases until the ratio of survivors to nonsurvivors is suitable. Final results will be presented at the conference.

ACKNOWLEDGMENTS

We would like to thank the Medical Research Council (MRC) of Canada and the Natural Science and Engineering Research Council (NSERC) of Canada for grants supporting this research. We also thank Richard E. Shaw, Michael B. Pliam, and Alex Zapolanski for use of their database.

REFERENCES

1. MB Pliam, RE Shaw, A Zapolanski. Comparative analysis of coronary surgery risk stratification models. J Invas Cardiol 9:203-222, 1997.
2. CM Ennett, M Frize. Investigation into the strengths and limitations of artificial neural networks: an application to an adult ICU patient database. Proc AMIA'98 998, 1998.

Appendix E: CCR and ASE Curves of Best-Performing Networks

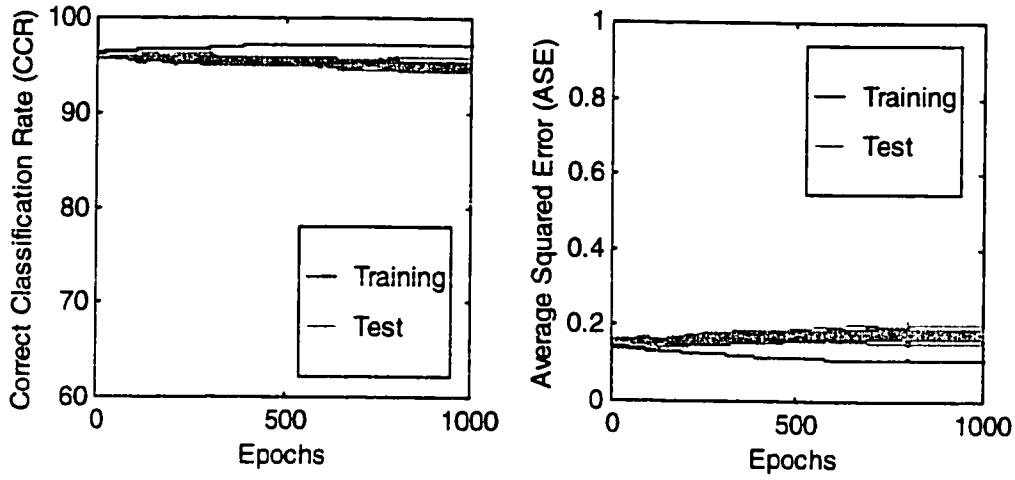


Figure E.1: CCR and ASE curves for an ANN with weight-elimination trained on the true mortality distribution (3.7 percent mortality) and tested on the true mortality distribution (3.8 percent mortality) (trorig/teorig)

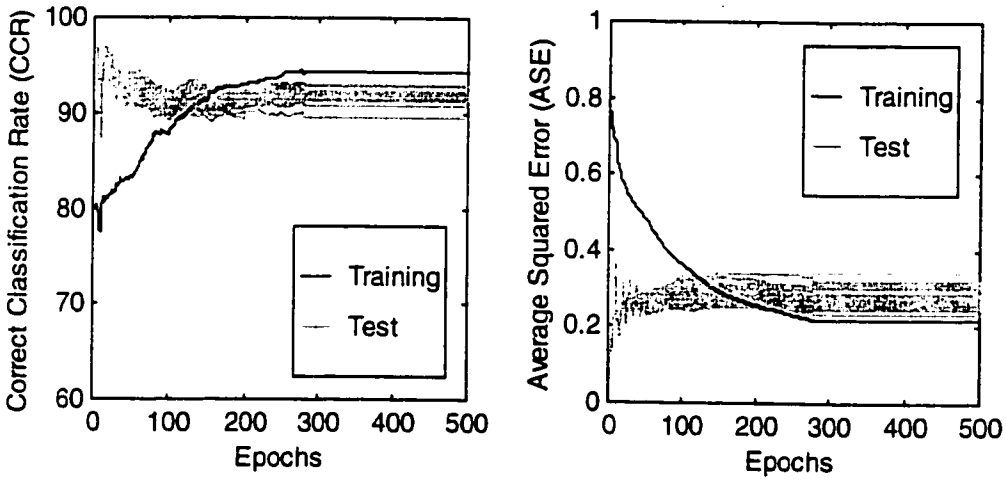


Figure E.2: CCR and ASE curves for an ANN with weight-elimination trained on an artificial dataset with a 20 percent mortality rate and tested on the true mortality distribution (tr20/teorig)

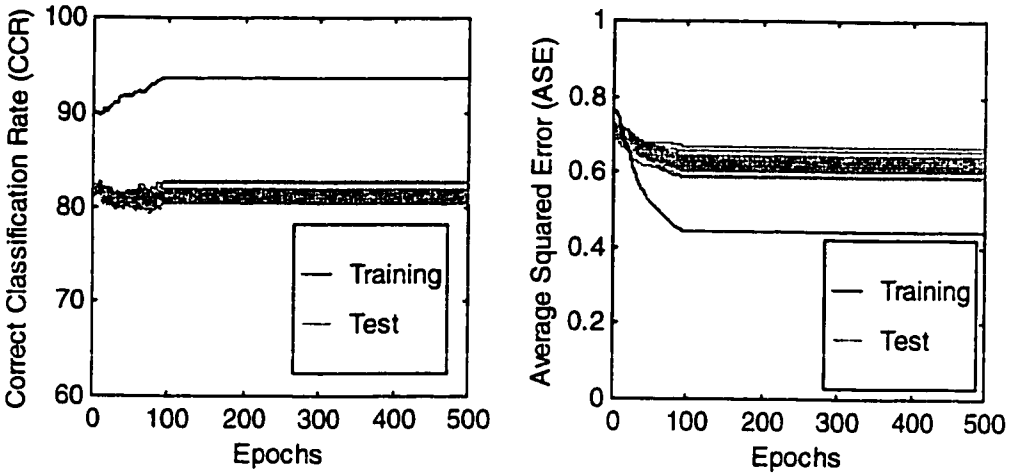


Figure E.3: CCR and ASE curves for an ANN with weight-elimination trained on an artificial dataset with a 10 percent mortality rate and tested on an artificial test set with a 20 percent mortality rate (tr10/te20)

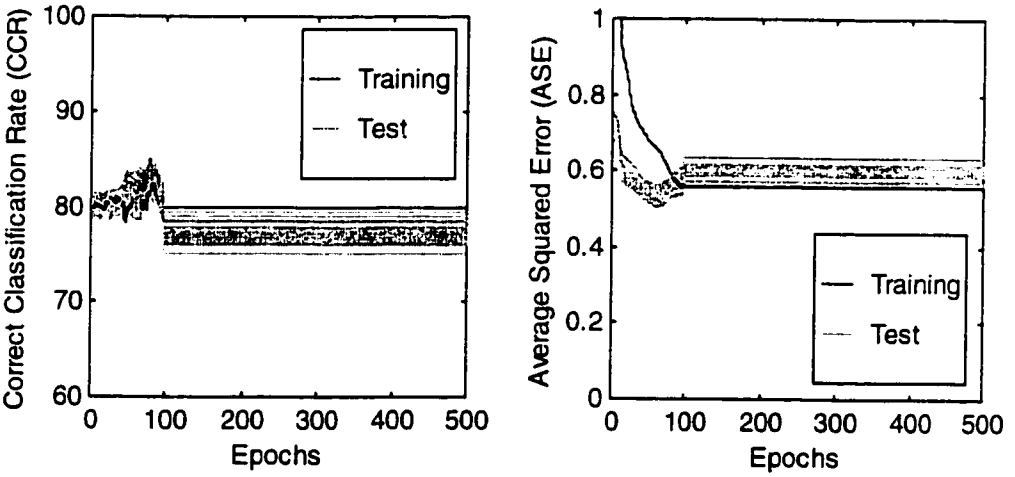


Figure E.4: CCR and ASE curves for an ANN with weight-elimination trained on an artificial dataset with a 20 percent mortality rate and tested on an artificial test set with a 20 percent mortality rate (tr20/te20)

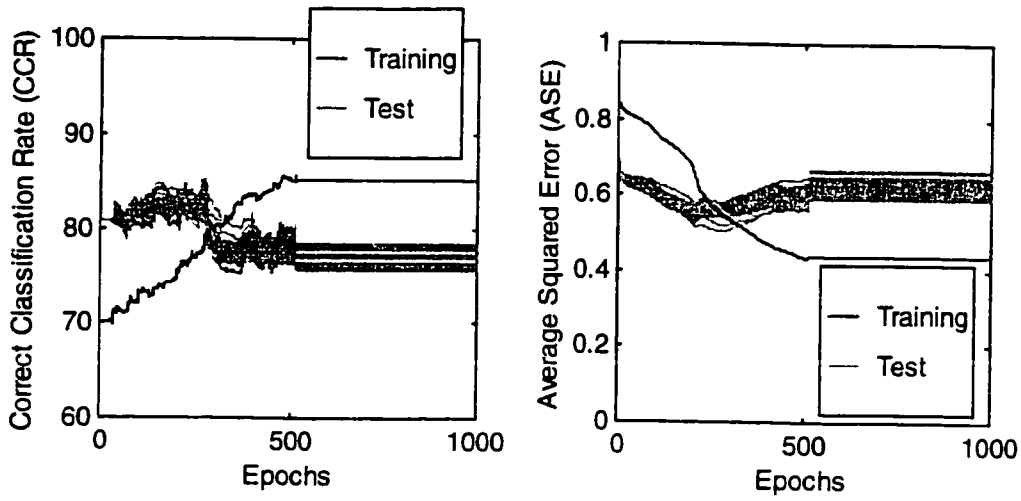


Figure E.5: CCR and ASE curves for an ANN with weight-elimination trained on artificial 30 percent mortality rate and tested on artificial test set with a 20 percent mortality rate (tr30/te20)

Appendix F: ROC Curves for Training and Test Sets

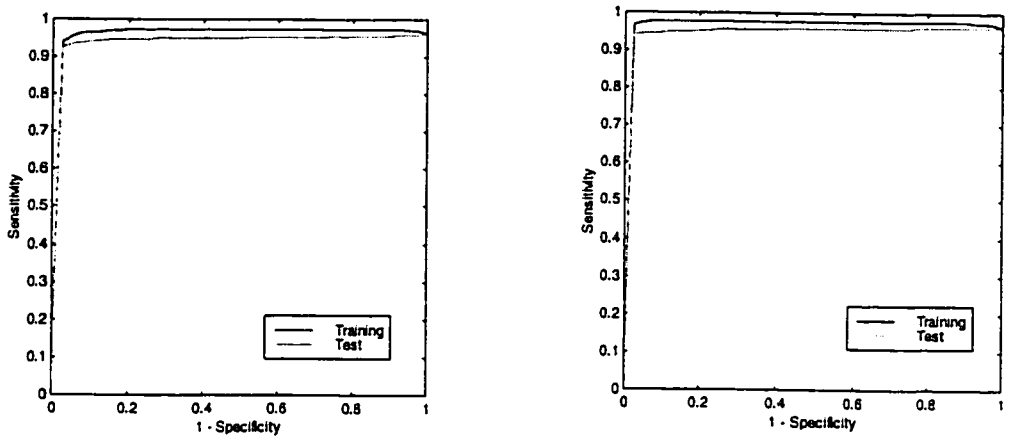


Figure F.1: ROC graphs at optimal sensitivity for an ANN trained and tested on the true mortality distributions (trorig/teorig) for (a) the weight-elimination network and (b) the ANN without weight-elimination

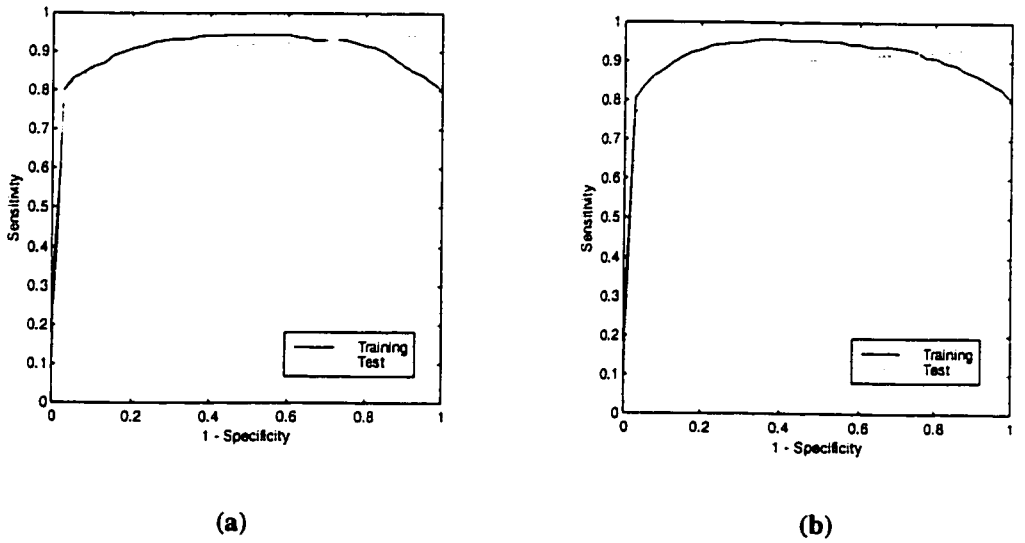
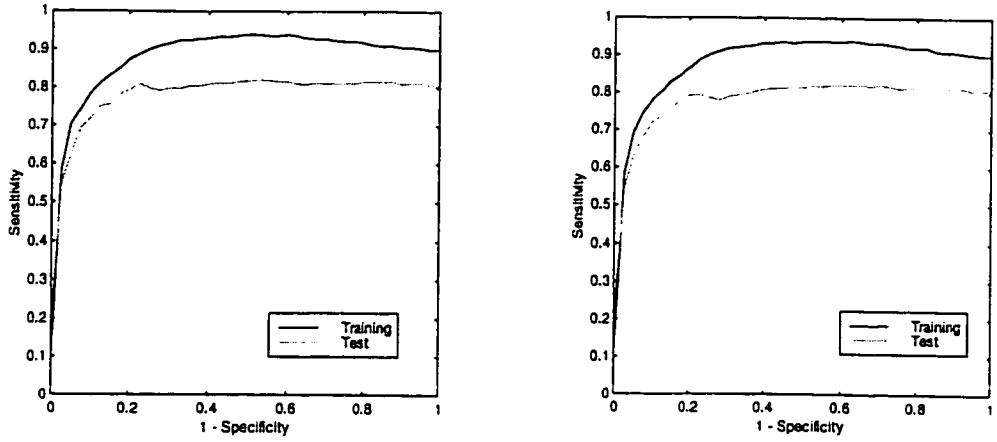
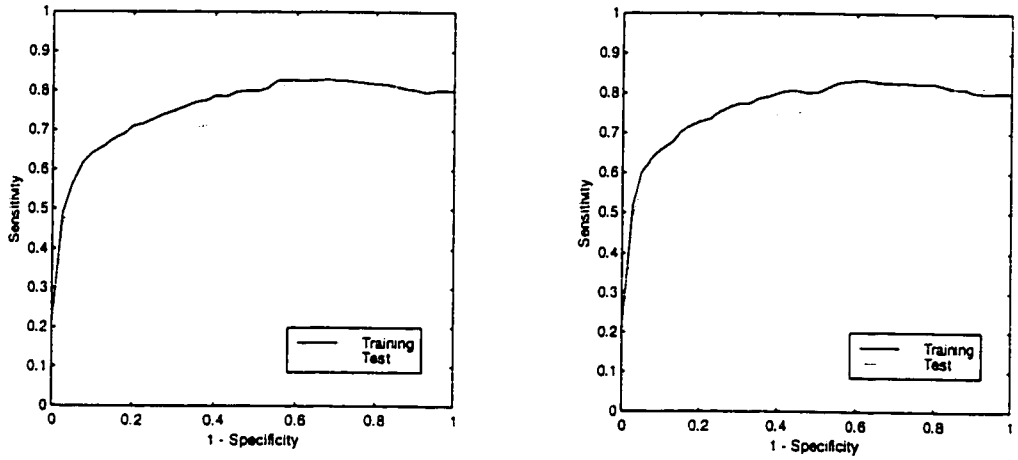


Figure F.2: ROC graphs at optimal sensitivity for an ANN trained on a dataset with a 20 percent mortality rate and tested on the true mortality distribution (tr20/teorig) for (a) the weight-elimination network and (b) the ANN without weight-elimination



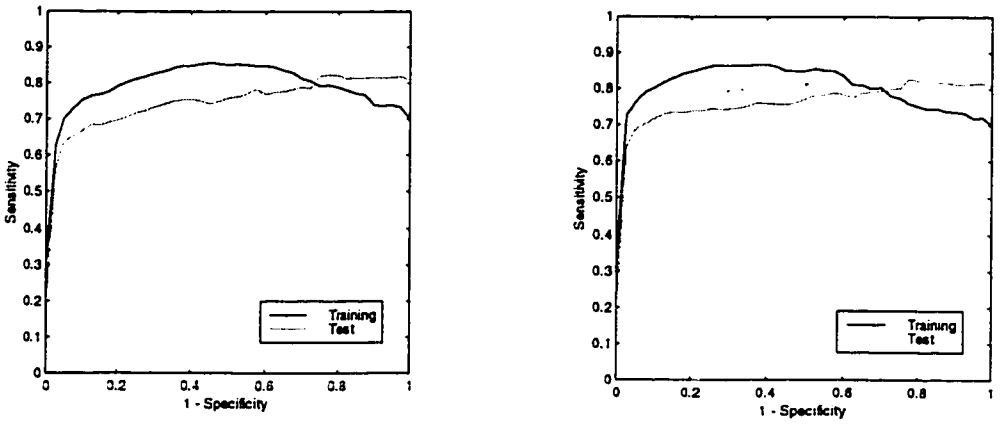
(a) (b)

Figure F.3: ROC graphs at optimal sensitivity for an ANN trained on a dataset with a 10 percent mortality rate and tested on a 20 percent mortality distribution (tr10/te20) for (a) the weight-elimination network and (b) the ANN without weight-elimination



(a) (b)

Figure F.4: ROC graphs at optimal sensitivity for an ANN trained on a dataset with a 20 percent mortality rate and tested on a 20 percent mortality distribution (tr20/te20) for (a) the weight-elimination network and (b) the ANN without weight-elimination



(a) (b)

Figure F.5: ROC graphs at optimal sensitivity for an ANN trained on a dataset with a 30 percent mortality rate and tested on a 20 percent mortality distribution (tr30/te20) for (a) the weight-elimination network and (b) the ANN without weight-elimination

Appendix G: ANN Performance on Different Test Sets

Table G.1: Performance on different test sets for an ANN with weight-elimination trained on the true mortality distribution (3.7 percent mortality) and tested on the true mortality distribution (3.8 percent mortality) (trorig/teorig)

Test set number	Area under the ROC curve	Mean predicted mortality	Sensitivity (/44)	Sensitivity	Specificity (/1129)	Specificity	PPV	PNV	CCR
original set	0.9391	0.0202	5	0.1136	1110	0.9832	0.2083	0.9661	0.9506
1	0.9382	0.016	1	0.0227	1113	0.9858	0.0588	0.9628	0.9497
2	0.9365	0.0273	8	0.1818	1103	0.9770	0.2353	0.9684	0.9471
3	0.9403	0.0183	4	0.0909	1112	0.9849	0.1905	0.9653	0.9514
4	0.9373	0.0178	3	0.0682	1108	0.9814	0.1250	0.9643	0.9471
5	0.9326	0.0277	5	0.1136	1102	0.9761	0.1563	0.9658	0.9437
6	0.9372	0.0183	2	0.0455	1110	0.9832	0.0952	0.9635	0.9480
7	0.9359	0.0217	4	0.0909	1104	0.9779	0.1379	0.9650	0.9446
8	0.9390	0.0179	4	0.0909	1112	0.9849	0.1905	0.9653	0.9514
9	0.9371	0.0233	5	0.1136	1106	0.9796	0.1786	0.9659	0.9471
10	0.9387	0.0221	5	0.1136	1108	0.9814	0.1923	0.9660	0.9488
11	0.9432	0.0213	8	0.1818	1113	0.9858	0.3333	0.9687	0.9557
12	0.9416	0.0159	4	0.0909	1113	0.9858	0.2000	0.9653	0.9523
13	0.9352	0.0303	9	0.2045	1101	0.9752	0.2432	0.9692	0.9463
14	0.9369	0.0260	7	0.1591	1105	0.9787	0.2258	0.9676	0.9480
15	0.9373	0.0174	2	0.0455	1110	0.9832	0.0952	0.9635	0.9480
16	0.9394	0.0206	7	0.1591	1110	0.9832	0.2692	0.9677	0.9523
17	0.9375	0.0203	3	0.0682	1110	0.9832	0.1364	0.9644	0.9488
18	0.9421	0.0141	2	0.0455	1116	0.9885	0.1333	0.9637	0.9531
19	0.9394	0.0203	6	0.1364	1108	0.9814	0.2222	0.9668	0.9497
20	0.9389	0.0198	5	0.1136	1110	0.9832	0.2083	0.9661	0.9506
21	0.9344	0.0225	3	0.0682	1103	0.9770	0.1034	0.9642	0.9429
22	0.9386	0.0223	7	0.1591	1107	0.9805	0.2414	0.9677	0.9497
23	0.9405	0.0174	4	0.0909	1113	0.9858	0.2000	0.9653	0.9523
24	0.9377	0.0199	4	0.0909	1110	0.9832	0.1739	0.9652	0.9497
25	0.9343	0.0262	5	0.1136	1103	0.9770	0.1613	0.9658	0.9446
26	0.9469	0.0144	7	0.1591	1118	0.9903	0.3889	0.9680	0.9591
27	0.9409	0.0193	5	0.1136	1112	0.9849	0.2273	0.9661	0.9523
28	0.9410	0.0169	4	0.0909	1114	0.9867	0.2105	0.9653	0.9531
29	0.9384	0.0216	5	0.1136	1109	0.9823	0.2000	0.9660	0.9497
30	0.9430	0.0163	5	0.1136	1116	0.9885	0.2778	0.9662	0.9557
mean	**0.9387**	**0.0204**	**4.7742**	**0.1085**	**1109.3226**	**0.9826**	**0.1942**	**0.9659**	**0.9498**
std dev	**0.0029**	**0.0040**	**1.9272**	**0.0438**	**4.3847**	**0.0039**	**0.0693**	**0.0016**	**0.0036**

Table G.2: Performance on different test sets for an ANN with weight-elimination trained on an artificial dataset with a 20 percent mortality rate and tested on the true mortality distribution (tr20/teorig)

Test set number	Area under the ROC curve	Mean predicted mortality	Sensitivity (/44)	Sensitivity	Specificity (/1129)	Specificity	PPV	PNV	CCR
original set	0.8963	0.0880	19	0.4318	1057	0.9362	0.2088	0.9769	0.9173
1	0.9068	0.0733	15	0.3409	1076	0.9531	0.2206	0.9738	0.9301
2	0.9051	0.0864	24	0.5455	1060	0.9389	0.2581	0.9815	0.9241
3	0.9045	0.0832	22	0.5000	1066	0.9442	0.2588	0.9798	0.9275
4	0.8983	0.0824	19	0.4318	1059	0.9380	0.2135	0.9769	0.9190
5	0.8874	0.0984	20	0.4545	1046	0.9265	0.1942	0.9776	0.9088
6	0.8971	0.0881	20	0.4545	1058	0.9371	0.2198	0.9778	0.9190
7	0.8804	0.0953	14	0.3182	1036	0.9176	0.1308	0.9719	0.8951
8	0.9006	0.0884	23	0.5227	1067	0.9451	0.2706	0.9807	0.9292
9	0.8999	0.0811	17	0.3864	1062	0.9407	0.2024	0.9752	0.9199
10	0.9039	0.0803	19	0.4318	1069	0.9469	0.2405	0.9771	0.9275
11	0.8978	0.0892	20	0.4545	1059	0.9380	0.2222	0.9778	0.9199
12	0.9040	0.0819	21	0.4773	1063	0.9415	0.2414	0.9788	0.9241
13	0.8928	0.0927	19	0.4318	1054	0.9336	0.2021	0.9768	0.9147
14	0.9046	0.0828	22	0.5000	1067	0.9451	0.2619	0.9798	0.9284
15	0.8905	0.0868	16	0.3636	1056	0.9353	0.1798	0.9742	0.9139
16	0.8969	0.0955	24	0.5455	1048	0.9283	0.2286	0.9813	0.9139
17	0.8869	0.0903	14	0.3182	1053	0.9327	0.1556	0.9723	0.9096
18	0.9027	0.0797	19	0.4318	1062	0.9407	0.2209	0.9770	0.9216
19	0.9083	0.0750	18	0.4091	1073	0.9504	0.2432	0.9763	0.9301
20	0.9096	0.0762	20	0.4545	1071	0.9486	0.2564	0.9781	0.9301
21	0.8916	0.0903	17	0.3864	1054	0.9336	0.1848	0.9750	0.9130
22	0.8881	0.1018	22	0.5000	1044	0.9247	0.2056	0.9794	0.9088
23	0.8992	0.0804	18	0.4091	1065	0.9433	0.2195	0.9762	0.9233
24	0.9025	0.0903	24	0.5455	1056	0.9353	0.2474	0.9814	0.9207
25	0.8987	0.0946	25	0.5682	1055	0.9345	0.2525	0.9823	0.9207
26	0.8900	0.0861	13	0.2955	1058	0.9371	0.1548	0.9715	0.9130
27	0.8975	0.0930	23	0.5227	1058	0.9371	0.2447	0.9805	0.9216
28	0.8934	0.0868	15	0.3409	1061	0.9398	0.1807	0.9734	0.9173
29	0.9030	0.0742	15	0.3409	1069	0.9469	0.2000	0.9736	0.9241
30	0.8947	0.0868	17	0.3864	1056	0.9353	0.1889	0.9751	0.9147
mean	**0.8978**	**0.0864**	**19.1613**	**0.4355**	**1059.2903**	**0.9383**	**0.2164**	**0.9771**	**0.9194**
std dev	**0.0070**	**0.0071**	**3.3376**	**0.0759**	**8.6764**	**0.0077**	**0.0346**	**0.0030**	**0.0079**

Table G.3: Performance on different test sets for an ANN with weight-elimination trained on an artificial dataset with a 10 percent mortality rate and tested on an artificial test set with a 20 percent mortality rate (tr10/te20)

Test set number	Area under the ROC curve	Mean predicted mortality	Sensitivity (/226)	Sensitivity	Specificity (/947)	Specificity	PPV	PNV	CCR
original set	0.7823	0.1059	38	0.1681	922	0.9736	0.6032	0.8306	0.8184
1	0.7818	0.1151	47	0.2080	911	0.9620	0.5663	0.8358	0.8167
2	0.7846	0.1129	45	0.1991	913	0.9641	0.5696	0.8346	0.8167
3	0.7791	0.1134	45	0.1991	908	0.9588	0.5357	0.8338	0.8124
4	0.7900	0.1146	52	0.2301	914	0.9652	0.6118	0.8401	0.8235
5	0.7812	0.1152	47	0.2080	908	0.9588	0.5465	0.8353	0.8142
6	0.7789	0.1065	40	0.1770	917	0.9683	0.5714	0.8314	0.8159
7	0.7834	0.1104	45	0.1991	906	0.9567	0.5233	0.8335	0.8107
8	0.7783	0.1093	43	0.1903	906	0.9567	0.5119	0.8320	0.8090
9	0.7839	0.1100	43	0.1903	912	0.9630	0.5513	0.8329	0.8142
10	0.7795	0.1053	36	0.1593	913	0.9641	0.5143	0.8277	0.8090
11	0.7849	0.1056	41	0.1814	918	0.9694	0.5857	0.8323	0.8176
12	0.7775	0.1106	40	0.1770	904	0.9546	0.4819	0.8294	0.8048
13	0.7836	0.1145	44	0.1947	909	0.9599	0.5366	0.8332	0.8124
14	0.7769	0.1170	42	0.1858	909	0.9599	0.5250	0.8317	0.8107
15	0.7896	0.1155	51	0.2257	918	0.9694	0.6375	0.8399	0.8261
16	0.7781	0.1114	39	0.1726	907	0.9578	0.4937	0.8291	0.8065
17	0.7788	0.1250	46	0.2035	905	0.9556	0.5227	0.8341	0.8107
18	0.7798	0.1095	39	0.1726	910	0.9609	0.5132	0.8295	0.8090
19	0.7732	0.1060	31	0.1372	912	0.9630	0.4697	0.8238	0.8039
20	0.7716	0.1072	32	0.1416	909	0.9599	0.4571	0.8241	0.8022
21	0.7778	0.1100	40	0.1770	915	0.9662	0.5556	0.8311	0.8142
22	0.7837	0.1138	42	0.1858	914	0.9652	0.5600	0.8324	0.8150
23	0.7770	0.1127	40	0.1770	910	0.9609	0.5195	0.8303	0.8099
24	0.7785	0.1083	39	0.1726	915	0.9662	0.5493	0.8303	0.8133
25	0.7791	0.1107	35	0.1549	916	0.9673	0.5303	0.8275	0.8107
26	0.7886	0.1097	49	0.2168	919	0.9704	0.6364	0.8385	0.8252
27	0.7837	0.1047	39	0.1726	920	0.9715	0.5909	0.8311	0.8176
28	0.7813	0.1030	37	0.1637	921	0.9725	0.5873	0.8297	0.8167
29	0.7819	0.1128	46	0.2035	913	0.9641	0.5750	0.8353	0.8176
30	0.7801	0.1046	35	0.1549	912	0.9630	0.5000	0.8268	0.8073
mean	**0.7809**	**0.1107**	**41.5484**	**0.1838**	**912.4516**	**0.9635**	**0.5462**	**0.8319**	**0.8133**
std dev	**0.0042**	**0.0046**	**5.1241**	**0.0227**	**4.8569**	**0.0051**	**0.0450**	**0.0039**	**0.0058**

Table G.4: Performance on different test sets for an ANN with weight-elimination trained on an artificial dataset with a 20 percent mortality rate and tested on an artificial test set with a 20 percent mortality rate (tr20/te20)

Test set number	Area under the ROC curve	Mean predicted mortality	Sensitivity (/226)	Sensitivity	Specificity (/947)	Specificity	PPV	PNV	CCR
original set	0.7321	0.2319	117	0.5177	795	0.8395	0.4349	0.8794	0.7775
1	0.7286	0.2366	124	0.5487	773	0.8163	0.4161	0.8834	0.7647
2	0.7292	0.2362	127	0.5619	782	0.8258	0.4349	0.8876	0.7749
3	0.7243	0.2398	121	0.5354	782	0.8258	0.4231	0.8816	0.7698
4	0.7266	0.2326	114	0.5044	784	0.8279	0.4116	0.8750	0.7656
5	0.7298	0.2380	117	0.5177	793	0.8374	0.4317	0.8792	0.7758
6	0.7263	0.2254	109	0.4823	789	0.8332	0.4082	0.8709	0.7656
7	0.7253	0.2317	121	0.5354	799	0.8437	0.4498	0.8838	0.7843
8	0.7251	0.2361	121	0.5354	772	0.8152	0.4088	0.8803	0.7613
9	0.7365	0.2328	125	0.5531	779	0.8226	0.4266	0.8852	0.7707
10	0.7365	0.2296	129	0.5708	801	0.8458	0.4691	0.8920	0.7928
11	0.7393	0.2279	123	0.5442	797	0.8416	0.4505	0.8856	0.7843
12	0.7293	0.2390	124	0.5487	782	0.8258	0.4291	0.8846	0.7724
13	0.7214	0.2347	112	0.4956	781	0.8247	0.4029	0.8726	0.7613
14	0.7209	0.2424	113	0.5000	766	0.8089	0.3844	0.8714	0.7494
15	0.7347	0.2442	136	0.6018	791	0.8353	0.4658	0.8978	0.7903
16	0.7293	0.2412	123	0.5442	767	0.8099	0.4059	0.8816	0.7587
17	0.7265	0.2424	126	0.5575	770	0.8131	0.4158	0.8851	0.7639
18	0.7284	0.2349	122	0.5398	788	0.8321	0.4342	0.8834	0.7758
19	0.7250	0.2419	117	0.5177	772	0.8152	0.4007	0.8763	0.7579
20	0.7309	0.2341	118	0.5221	794	0.8384	0.4354	0.8803	0.7775
21	0.7235	0.2440	124	0.5487	783	0.8268	0.4306	0.8847	0.7732
22	0.7351	0.2288	125	0.5531	783	0.8268	0.4325	0.8857	0.7741
23	0.7233	0.2442	126	0.5575	794	0.8384	0.4516	0.8881	0.7843
24	0.7394	0.2261	118	0.5221	801	0.8458	0.4470	0.8812	0.7835
25	0.7266	0.2379	124	0.5487	775	0.8184	0.4189	0.8837	0.7664
26	0.7287	0.2395	124	0.5487	794	0.8384	0.4477	0.8862	0.7826
27	0.7298	0.2289	114	0.5044	792	0.8363	0.4238	0.8761	0.7724
28	0.7295	0.2350	111	0.4912	782	0.8258	0.4022	0.8718	0.7613
29	0.7196	0.2389	114	0.5044	777	0.8205	0.4014	0.8740	0.7596
30	0.7229	0.2326	110	0.4867	787	0.8310	0.4074	0.8715	0.7647
mean	**0.7285**	**0.2358**	**120.2903**	**0.5323**	**784.6774**	**0.8286**	**0.4259**	**0.8813**	**0.7715**
std dev	**0.0052**	**0.0054**	**6.2407**	**0.0276**	**10.0379**	**0.0106**	**0.0204**	**0.0064**	**0.0104**

Table G.5: Performance on different test sets for an ANN with weight-elimination trained on artificial 30 percent mortality rate and tested on artificial test set with a 20 percent mortality rate (tr30/te20)

Test set number	Area under the ROC curve	Mean predicted mortality	Sensitivity (/226)	Sensitivity	Specificity (/947)	Specificity	PPV	PNV	CCR
original set	0.7460	0.2535	134	0.5929	757	0.7994	0.4136	0.8916	0.7596
1	0.7549	0.2593	148	0.6549	754	0.7962	0.4340	0.9063	0.7690
2	0.7531	0.2539	138	0.6106	757	0.7994	0.4207	0.8959	0.7630
3	0.7551	0.2639	154	0.6814	752	0.7941	0.4413	0.9126	0.7724
4	0.7538	0.2583	148	0.6549	755	0.7973	0.4353	0.9064	0.7698
5	0.7496	0.2567	143	0.6327	758	0.8004	0.4307	0.9013	0.7681
6	0.7526	0.2498	141	0.6239	758	0.8004	0.4273	0.8992	0.7664
7	0.7510	0.2597	137	0.6062	751	0.7930	0.4114	0.8940	0.7570
8	0.7537	0.2764	158	0.6991	737	0.7782	0.4293	0.9155	0.7630
9	0.7643	0.2595	151	0.6681	761	0.8036	0.4481	0.9103	0.7775
10	0.7606	0.2392	127	0.5619	776	0.8194	0.4262	0.8869	0.7698
11	0.7630	0.2533	152	0.0011	762	0.8046	0.4510	0.9115	0.7792
12	0.7515	0.2563	141	0.6239	752	0.7941	0.4196	0.8984	0.7613
13	0.7422	0.2642	141	0.6239	748	0.7899	0.4147	0.8980	0.7579
14	0.7486	0.2609	146	0.6460	746	0.7878	0.4207	0.9031	0.7604
15	0.7599	0.2577	152	0.6726	764	0.8068	0.4537	0.9117	0.7809
16	0.7526	0.2729	158	0.6991	745	0.7867	0.4389	0.9164	0.7698
17	0.7475	0.2644	139	0.6150	745	0.7867	0.4076	0.8954	0.7536
18	0.7555	0.2623	149	0.6593	753	0.7951	0.4344	0.9072	0.7690
19	0.7593	0.2552	142	0.6283	764	0.8068	0.4369	0.9009	0.7724
20	0.7494	0.2504	138	0.6106	763	0.8057	0.4286	0.8966	0.7681
21	0.7455	0.2625	141	0.6239	749	0.7909	0.4159	0.8981	0.7587
22	0.7468	0.2505	135	0.5973	754	0.7962	0.4116	0.8923	0.7579
23	0.7482	0.2593	140	0.6195	750	0.7920	0.4154	0.8971	0.7587
24	0.7635	0.2469	149	0.6593	768	0.8110	0.4543	0.9089	0.7818
25	0.7577	0.2639	154	0.6814	756	0.7983	0.4464	0.9130	0.7758
26	0.7600	0.2590	147	0.6504	759	0.8015	0.4388	0.9057	0.7724
27	0.7539	0.2551	142	0.6283	758	0.8004	0.4290	0.9002	0.7673
28	0.7526	0.2553	145	0.6416	757	0.7994	0.4328	0.9033	0.7690
29	0.7580	0.2541	144	0.6372	762	0.8046	0.4377	0.9028	0.7724
30	0.7474	0.2512	139	0.6150	762	0.8046	0.4290	0.8975	0.7681
mean	**0.7535**	**0.2576**	**144.2903**	**0.6168**	**755.9032**	**0.7982**	**0.4302**	**0.9025**	**0.7674**
std dev	**0.0057**	**0.0072**	**7.2212**	**0.1185**	**7.7561**	**0.0082**	**0.0130**	**0.0075**	**0.0074**

Table G.6: Performance on different test sets for an ANN without weight-elimination trained on the true mortality distribution (3.7 percent mortality) and tested on the true mortality distribution (3.8 percent mortality) (trorig/teorig)

Test set number	Area under the ROC curve	Mean predicted mortality	Sensitivity (/226)	Sensitivity	Specificity (/947)	Specificity	PPV	PNV	CCR
original set	0.9452	0.0121	4	0.0909	1121	0.9929	0.3333	0.9655	0.9591
1	0.9417	0.0130	2	0.0455	1116	0.9885	0.1333	0.9637	0.9531
2	0.9464	0.0108	4	0.0909	1123	0.9947	0.4000	0.9656	0.9608
3	0.9462	0.0071	1	0.0227	1126	0.9973	0.2500	0.9632	0.9608
4	0.9486	0.0086	5	0.1136	1126	0.9973	0.6250	0.9665	0.9642
5	0.9457	0.0108	3	0.0682	1123	0.9947	0.3333	0.9648	0.9599
6	0.9442	0.0108	2	0.0455	1122	0.9938	0.2222	0.9639	0.9582
7	0.9418	0.0125	2	0.0455	1119	0.9911	0.1667	0.9638	0.9557
8	0.9429	0.0121	3	0.0682	1119	0.9911	0.2308	0.9647	0.9565
9	0.9448	0.0110	3	0.0682	1120	0.9920	0.2500	0.9647	0.9574
10	0.9454	0.0111	3	0.0682	1122	0.9938	0.3000	0.9647	0.9591
11	0.9465	0.0122	5	0.1136	1121	0.9929	0.3846	0.9664	0.9599
12	0.9493	0.0070	3	0.0682	1125	0.9965	0.4286	0.9648	0.9616
13	0.9458	0.0173	8	0.1818	1118	0.9903	0.4211	0.9688	0.9599
14	0.9469	0.0138	6	0.1364	1120	0.9920	0.4000	0.9672	0.9599
15	0.9459	0.0103	3	0.0682	1123	0.9947	0.3333	0.9648	0.9599
16	0.9450	0.0147	7	0.1591	1119	0.9911	0.4118	0.9680	0.9599
17	0.9419	0.0130	2	0.0455	1120	0.9920	0.1818	0.9639	0.9565
18	0.9446	0.0117	3	0.0682	1120	0.9920	0.2500	0.9647	0.9574
19	0.9488	0.0125	7	0.1591	1123	0.9947	0.5385	0.9681	0.9633
20	0.9476	0.0094	4	0.0909	1124	0.9956	0.4444	0.9656	0.9616
21	0.9421	0.0116	1	0.0227	1121	0.9929	0.1111	0.9631	0.9565
22	0.9462	0.0128	6	0.1364	1122	0.9938	0.4615	0.9672	0.9616
23	0.9425	0.0099	1	0.0227	1121	0.9929	0.1111	0.9631	0.9565
24	0.9458	0.0124	5	0.1136	1120	0.9920	0.3571	0.9664	0.9591
25	0.9450	0.0117	3	0.0682	1121	0.9929	0.2727	0.9647	0.9582
26	0.9495	0.0147	9	0.2045	1121	0.9929	0.5294	0.9697	0.9633
27	0.9409	0.0174	4	0.0909	1114	0.9867	0.2105	0.9653	0.9531
28	0.9425	0.0130	3	0.0682	1119	0.9911	0.2308	0.9647	0.9565
29	0.9480	0.0117	5	0.1136	1124	0.9956	0.5000	0.9665	0.9625
30	0.9466	0.0123	5	0.1136	1123	0.9947	0.4545	0.9664	0.9616
mean	**0.9453**	**0.0119**	**3.9355**	**0.0894**	**1121.1613**	**0.9931**	**0.3315**	**0.9655**	**0.9592**
std dev	**0.0023**	**0.0023**	**1.9990**	**0.0454**	**2.6039**	**0.0023**	**0.1314**	**0.0017**	**0.0027**

Table G.7: Performance on different test sets for an ANN without weight-elimination trained on an artificial dataset with a 20 percent mortality rate and tested on the true mortality distribution (tr20/teorig)

Test set number	Area under the ROC curve	Mean predicted mortality	Sensitivity (/44)	Sensitivity	Specificity (/1129)	Specificity	PPV	PNV	CCR
original set	0.8833	0.0936	11	0.2500	1051	0.9309	0.1236	0.9696	0.9054
1	0.8940	0.0828	10	0.2273	1065	0.9433	0.1351	0.9691	0.9165
2	0.8969	0.0874	17	0.3864	1063	0.9415	0.2048	0.9752	0.9207
3	0.8856	0.0980	15	0.3409	1045	0.9256	0.1515	0.9730	0.9037
4	0.8839	0.0857	6	0.1364	1057	0.9362	0.0769	0.9653	0.9062
5	0.8713	0.1038	10	0.2273	1041	0.9221	0.1020	0.9684	0.8960
6	0.8921	0.0871	12	0.2727	1061	0.9398	0.1500	0.9707	0.9147
7	0.8751	0.0947	7	0.1591	1043	0.9238	0.0753	0.9657	0.8951
8	0.8887	0.0938	14	0.3182	1057	0.9362	0.1628	0.9724	0.9130
9	0.8852	0.0902	10	0.2273	1055	0.9345	0.1190	0.9688	0.9079
10	0.8902	0.0890	12	0.2727	1058	0.9371	0.1446	0.9706	0.9122
11	0.8880	0.0899	12	0.2727	1055	0.9345	0.1395	0.9706	0.9096
12	0.8847	0.0945	13	0.2955	1047	0.9274	0.1368	0.9712	0.9037
13	0.8874	0.0887	9	0.2045	1060	0.9389	0.1154	0.9680	0.9113
14	0.8868	0.0913	12	0.2727	1045	0.9256	0.1250	0.9703	0.9011
15	0.8766	0.0859	2	0.0455	1051	0.9309	0.0250	0.9616	0.8977
16	0.8770	0.1012	11	0.2500	1042	0.9229	0.1122	0.9693	0.8977
17	0.8803	0.0925	8	0.1818	1055	0.9345	0.0976	0.9670	0.9062
18	0.8817	0.0922	9	0.2045	1047	0.9274	0.0989	0.9677	0.9003
19	0.8975	0.0807	11	0.2500	1065	0.9433	0.1467	0.9699	0.9173
20	0.8879	0.0903	11	0.2500	1054	0.9336	0.1279	0.9696	0.9079
21	0.8757	0.1041	12	0.2727	1043	0.9238	0.1224	0.9702	0.8994
22	0.8774	0.1023	13	0.2955	1047	0.9274	0.1368	0.9712	0.9037
23	0.8896	0.0758	4	0.0909	1067	0.9451	0.0606	0.9639	0.9130
24	0.8831	0.1018	15	0.3409	1043	0.9238	0.1485	0.9729	0.9020
25	0.8899	0.0937	14	0.3182	1052	0.9318	0.1538	0.9723	0.9088
26	0.8708	0.1026	9	0.2045	1041	0.9221	0.0928	0.9675	0.8951
27	0.8912	0.0943	17	0.3864	1059	0.9380	0.1954	0.9751	0.9173
28	0.8813	0.0916	8	0.1818	1051	0.9309	0.0930	0.9669	0.9028
29	0.8762	0.0922	5	0.1136	1053	0.9327	0.0617	0.9643	0.9020
30	0.8868	0.0871	9	0.2045	1054	0.9336	0.1071	0.9679	0.9062
mean	**0.8844**	**0.0922**	**10.5806**	**0.2405**	**1052.4839**	**0.9322**	**0.1207**	**0.9692**	**0.9063**
std dev	**0.0069**	**0.0067**	**3.4713**	**0.0789**	**7.5215**	**0.0067**	**0.0376**	**0.0031**	**0.0069**

Table G.8: Performance on different test sets for an ANN without weight-elimination trained on an artificial dataset with a 10 percent mortality rate and tested on an artificial test set with a 20 percent mortality rate (tr10/te20)

Test set number	Area under the ROC curve	Mean predicted mortality	Sensitivity (/226)	Sensitivity	Specificity (/947)	Specificity	PPV	PNV	CCR
0	0.7816	0.1096	38	0.1681	919	0.9704	0.5758	0.8302	0.8159
1	0.7836	0.1192	47	0.2080	911	0.9620	0.5663	0.8358	0.8167
2	0.7846	0.1174	45	0.1991	909	0.9599	0.5422	0.8339	0.8133
3	0.7807	0.1173	45	0.1991	908	0.9588	0.5357	0.8338	0.8124
4	0.7907	0.1180	52	0.2301	915	0.9662	0.6190	0.8402	0.8244
5	0.7824	0.1191	47	0.2080	905	0.9556	0.5281	0.8349	0.8116
6	0.7799	0.1098	40	0.1770	917	0.9683	0.5714	0.8314	0.8159
7	0.7854	0.1121	45	0.1991	906	0.9567	0.5233	0.8335	0.8107
8	0.7805	0.1124	43	0.1903	904	0.9546	0.5000	0.8316	0.8073
9	0.7834	0.1148	43	0.1903	905	0.9556	0.5059	0.8318	0.8082
10	0.7807	0.1079	36	0.1593	911	0.9620	0.5000	0.8274	0.8073
11	0.7864	0.1091	41	0.1814	915	0.9662	0.5616	0.8318	0.8150
12	0.7789	0.1133	40	0.1770	909	0.9599	0.5128	0.8301	0.8090
13	0.7830	0.1176	44	0.1947	906	0.9567	0.5176	0.8327	0.8099
14	0.7782	0.1175	42	0.1858	908	0.9588	0.5185	0.8315	0.8099
15	0.7892	0.1182	51	0.2257	913	0.9641	0.6000	0.8392	0.8218
16	0.7783	0.1153	39	0.1726	908	0.9588	0.5000	0.8292	0.8073
17	0.7797	0.1267	46	0.2035	905	0.9556	0.5227	0.8341	0.8107
18	0.7818	0.1116	39	0.1726	915	0.9662	0.5493	0.8303	0.8133
19	0.7728	0.1091	31	0.1372	911	0.9620	0.4627	0.8237	0.8031
20	0.7729	0.1080	32	0.1416	908	0.9588	0.4507	0.8240	0.8014
21	0.7783	0.1135	40	0.1770	914	0.9652	0.5479	0.8309	0.8133
22	0.7830	0.1175	42	0.1858	913	0.9641	0.5526	0.8323	0.8142
23	0.7778	0.1158	40	0.1770	908	0.9588	0.5063	0.8300	0.8082
24	0.7793	0.1127	39	0.1726	917	0.9683	0.5652	0.8306	0.8150
25	0.7797	0.1128	35	0.1549	916	0.9673	0.5303	0.8275	0.8107
26	0.7884	0.1144	49	0.2168	916	0.9673	0.6125	0.8381	0.8227
27	0.7834	0.1078	39	0.1726	918	0.9694	0.5735	0.8308	0.8159
28	0.7817	0.1067	37	0.1637	920	0.9715	0.5781	0.8296	0.8159
29	0.7830	0.1157	46	0.2035	910	0.9609	0.5542	0.8349	0.8150
30	0.7803	0.1079	35	0.1549	909	0.9599	0.4795	0.8264	0.8048
mean	**0.7816**	**0.1138**	**41.5484**	**0.1838**	**911.2581**	**0.9623**	**0.5375**	**0.8317**	**0.8123**
std dev	**0.0039**	**0.0045**	**5.0407**	**0.0223**	**4.5576**	**0.0048**	**0.0399**	**0.0038**	**0.0052**

Table G.9: Performance on different test sets for an ANN without weight-elimination trained on an artificial dataset with a 20 percent mortality rate and tested on an artificial test set with a 20 percent mortality rate (tr20/te20)

Test set number	Area under the ROC curve	Mean predicted mortality	Sensitivity (/226)	Sensitivity	Specificity (/947)	Specificity	PPV	PNV	CCR
original set	0.7410	0.2195	94	0.4159	809	0.8543	0.4052	0.8597	0.7698
1	0.7403	0.2220	104	0.4602	806	0.8511	0.4245	0.8685	0.7758
2	0.7381	0.2246	105	0.4646	809	0.8543	0.4321	0.8699	0.7792
3	0.7341	0.2264	96	0.4248	813	0.8585	0.4174	0.8621	0.7749
4	0.7362	0.2214	92	0.4071	807	0.8522	0.3966	0.8576	0.7664
5	0.7394	0.2249	97	0.4292	808	0.8532	0.4110	0.8623	0.7715
6	0.7362	0.2117	83	0.3673	816	0.8617	0.3879	0.8509	0.7664
7	0.7350	0.2171	92	0.4071	823	0.8691	0.4259	0.8600	0.7801
8	0.7359	0.2238	95	0.4204	801	0.8458	0.3942	0.8594	0.7639
9	0.7452	0.2200	99	0.4381	805	0.8501	0.4108	0.8637	0.7707
10	0.7482	0.2146	100	0.4425	822	0.8680	0.4444	0.8671	0.7860
11	0.7484	0.2167	99	0.4381	824	0.8701	0.4459	0.8665	0.7869
12	0.7369	0.2251	96	0.4248	804	0.8490	0.4017	0.8608	0.7673
13	0.7323	0.2210	93	0.4115	808	0.8532	0.4009	0.8587	0.7681
14	0.7299	0.2289	93	0.4115	788	0.8321	0.3690	0.8556	0.7511
15	0.7455	0.2280	106	0.4690	815	0.8606	0.4454	0.8717	0.7852
16	0.7399	0.2287	104	0.4602	810	0.8553	0.4315	0.8691	0.7792
17	0.7363	0.2295	102	0.4513	796	0.8405	0.4032	0.8652	0.7656
18	0.7380	0.2215	101	0.4469	820	0.8659	0.4430	0.8677	0.7852
19	0.7333	0.2277	97	0.4292	791	0.8353	0.3834	0.8598	0.7570
20	0.7399	0.2206	86	0.3805	812	0.8574	0.3891	0.8529	0.7656
21	0.7341	0.2282	96	0.4248	801	0.8458	0.3967	0.8604	0.7647
22	0.7444	0.2151	95	0.4204	806	0.8511	0.4025	0.8602	0.7681
23	0.7337	0.2304	93	0.4115	851	0.8986	0.4921	0.8648	0.8048
24	0.7476	0.2120	92	0.4071	822	0.8680	0.4240	0.8598	0.7792
25	0.7379	0.2225	100	0.4425	808	0.8532	0.4184	0.8651	0.7741
26	0.7399	0.2244	93	0.4115	812	0.8574	0.4079	0.8593	0.7715
27	0.7390	0.2148	87	0.3850	819	0.8648	0.4047	0.8549	0.7724
28	0.7404	0.2188	97	0.4292	800	0.8448	0.3975	0.8611	0.7647
29	0.7312	0.2252	89	0.3938	802	0.8469	0.3803	0.8541	0.7596
30	0.7307	0.2200	88	0.3894	814	0.8596	0.3982	0.8550	0.7690
mean	**0.7384**	**0.2221**	**95.6129**	**0.4231**	**810.3871**	**0.8557**	**0.4124**	**0.8614**	**0.7724**
std dev	**0.0050**	**0.0052**	**5.5977**	**0.0248**	**11.4826**	**0.0121**	**0.0244**	**0.0052**	**0.0103**

Table G.10: Performance on different test sets for an ANN without weight-elimination trained on artificial 30 percent mortality rate and tested on artificial test set with ä 20 percent mortality rate (tr30/te20)

Test set number	Area under the ROC curve	Mean predicted mortality	Sensitivity (/226)	Sensitivity	Specificity (/947)	Specificity	PPV	PNV	CCR
original set	0.7621	0.2220	125	0.5531	784	0.8279	0.4340	0.8859	0.7749
1	0.7713	0.2276	138	0.6106	785	0.8289	0.4600	0.8992	0.7869
2	0.7723	0.2239	130	0.5752	783	0.8268	0.4422	0.8908	0.7783
3	0.7703	0.2320	143	0.6327	783	0.8268	0.4658	0.9042	0.7894
4	0.7704	0.2249	139	0.6150	795	0.8395	0.4777	0.9014	0.7962
5	0.7638	0.2304	133	0.5885	772	0.8152	0.4318	0.8925	0.7715
6	0.7667	0.2161	127	0.5619	789	0.8332	0.4456	0.8885	0.7809
7	0.7640	0.2240	123	0.5442	786	0.8300	0.4331	0.8841	0.7749
8	0.7662	0.2429	147	0.6504	764	0.8068	0.4455	0.9063	0.7766
9	0.7743	0.2295	141	0.6239	789	0.8332	0.4716	0.9027	0.7928
10	0.7743	0.2072	123	0.5442	801	0.8458	0.4572	0.8861	0.7877
11	0.7794	0.2189	141	0.6239	797	0.8416	0.4845	0.9036	0.7997
12	0.7651	0.2226	127	0.5619	784	0.8279	0.4379	0.8879	0.7766
13	0.7599	0.2333	135	0.5973	778	0.8215	0.4441	0.8953	0.7783
14	0.7632	0.2323	134	0.5929	780	0.8237	0.4452	0.8945	0.7792
15	0.7763	0.2283	140	0.6195	793	0.8374	0.4762	0.9022	0.7954
16	0.7677	0.2445	149	0.6593	768	0.8110	0.4543	0.9089	0.7818
17	0.7653	0.2314	130	0.5752	772	0.8152	0.4262	0.8894	0.7690
18	0.7727	0.2283	136	0.6018	788	0.8321	0.4610	0.8975	0.7877
19	0.7736	0.2218	134	0.5929	793	0.8374	0.4653	0.8960	0.7903
20	0.7625	0.2189	124	0.5487	783	0.8268	0.4306	0.8847	0.7732
21	0.7582	0.2296	127	0.5619	773	0.8163	0.4219	0.8865	0.7673
22	0.7612	0.2209	117	0.5177	764	0.8068	0.3900	0.8751	0.7511
23	0.7643	0.2261	131	0.5796	779	0.8226	0.4381	0.8913	0.7758
24	0.7794	0.2173	141	0.6239	789	0.8332	0.4716	0.9027	0.7928
25	0.7697	0.2342	140	0.6195	781	0.8247	0.4575	0.9008	0.7852
26	0.7778	0.2250	141	0.6239	796	0.8405	0.4829	0.9035	0.7988
27	0.7712	0.2188	130	0.5752	798	0.8427	0.4659	0.8926	0.7911
28	0.7660	0.2271	135	0.5973	779	0.8226	0.4455	0.8954	0.7792
29	0.7750	0.2200	138	0.6106	792	0.8363	0.4710	0.9000	0.7928
30	0.7663	0.2209	133	0.5885	788	0.8321	0.4555	0.8944	0.7852
mean	**0.7687**	**0.2258**	**133.9355**	**0.5926**	**784.0645**	**0.8279**	**0.4513**	**0.8950**	**0.7826**
std dev	**0.0057**	**0.0075**	**7.4571**	**0.0330**	**9.5645**	**0.0101**	**0.0203**	**0.0078**	**0.0105**

Appendix H: Variable Weights Following Weight-Elimination Approach

Training set: true mortality rate (trorig)
Test set: true mortality rate (teorig)

confu =

2171	61
0	22

confut =

1110	39
19	5

float = 0
layers = 2
h1 = 9
wtfact = 1

Final learning rate value
lr = 0.0109
lr_inc = 1.0030
lr_dec = 1

Final weight decay constant
lambda = 3.0000e-004
wnot = 0.1500
momentum = 0.7500
err_ratio = 1.0200
sens = 0.1136
spec = 0.9832
PPV = 0.2083
PNV = 0.9661
CCR = 0.9506

Training Set Apriori Probabilities
trprob =
96.3177
3.6823

Test Set Apriori Probabilities
teprob =
96.2489
3.7511

MDC Training Set Classification Rate
bayes_train = 96.2733

MDC Test Set Classification Rate
bayes_test = 95.8227

W1 =

Columns 1 through 7

0.0000	0.0331	-0.0116	0.0044	0.5575	-0.0041	-0.8898
-0.0046	-0.0139	0.0138	0.0006	-0.6390	-0.0163	0.9449
0.0100	-0.0290	0.0195	-1.3403	-0.0308	0.8329	0.0107
0.4068	0.0138	-0.6760	0.0053	-0.9873	1.2799	0.0011
0.0405	0.4096	-0.4863	-0.0085	1.2485	0.1473	-0.0123
-0.0005	0.7423	-0.0006	-0.4738	-0.0015	-0.8471	-0.0005
-1.1421	1.2276	-0.3459	-0.0138	0.8034	-0.0016	-0.2911
0.0004	-0.8295	0.0007	0.0004	0.0009	-0.6739	0.0010

```
    0.0169   -1.2202   -0.4069   -0.0113   -0.0463   -0.7214    0.5457
  Columns 8 through 14
   -0.0254    0.8054   -0.2355    0.9866   -0.0360   -0.0068    1.2683
    0.0278    0.8151   -0.0007    0.4976    0.0111   -0.4984   -0.1256
    0.0323    0.0020    0.6564    0.0157    0.7910   -0.0026   -0.9434
    1.4684    0.0038    0.6880    0.0121   -0.0113    0.3898    0.0037
   -0.0049   -0.0112   -0.0054    1.0414    1.5506   -0.0180   -0.0115
    0.0001   -0.0005   -0.0003   -0.0003    0.0008   -0.0006   -0.0015
    1.1808    0.3533    0.6919   -0.5853    0.0669   -0.8443   -1.1063
   -0.6563    0.0004    0.0003   -0.0008   -0.0001    0.0006    0.0008
   -0.6757   -0.7752    0.0017   -0.0109    0.0328   -0.0082   -0.0159
  Columns 15 through 21
    0.0066    0.0330    1.0005   -0.0037    0.0247    0.0246   -0.0451
    0.0019   -0.0034    0.8446   -0.0025    0.9363    0.0023    0.9134
   -0.0036   -1.3999   -0.0042   -1.0675   -0.0200   -0.0458    0.4159
    0.0855    0.0023    0.0100   -0.0003   -0.0027    0.6919    0.5595
   -0.0063   -0.0018    0.0114   -0.5524    0.3073    0.0116    0.6594
   -0.0004   -0.0004    0.0003    0.0005    0.5254   -0.0004   -0.0010
   -0.0054   -0.6035    0.0745    0.0110    0.7974    0.0132    0.0216
    0.0004   -0.0008    0.9586   -0.9079    0.9212   -0.8416    0.6864
   -0.0039    0.0139    1.1872    1.2127    0.0341    0.0784   -0.3413
  Columns 22 through 28
   -0.0253   -0.8321    0.7231   -0.6542    0.6431   -0.0346    0.0697
    0.0100    0.0285    0.0073   -0.4624    0.0018    0.0011    0.9462
    0.7718    0.5404    1.0304   -0.0036    0.5482    0.6972    1.0644
    0.0393    1.1123   -0.0109    0.0034    0.0031    0.0642    0.8490
    0.0016   -0.0226    1.2659    0.6332   -0.0062    0.0232   -0.2692
   -0.0004   -0.0002    0.0003   -0.0004   -0.0004    0.9893   -0.0004
   -0.0148    0.0148    0.0137   -0.0174   -0.0043    0.0294    0.0083
    0.0007    0.0011    0.0007    0.3090    0.0004    0.0005    0.8055
    1.2196   -0.6496   -0.8990   -0.8155   -0.7004    0.4487    1.2392
  Columns 29 through 32
   -0.6269    0.7633   -0.4386    0.0025
    0.0086   -0.0164   -0.0033    0.0086
   -0.6608    0.0293    0.0014    0.0287
    0.4719    0.0073    0.0008   -0.0202
    0.9227   -0.8919   -0.0018   -0.0159
   -0.0003   -0.0007   -0.0004   -0.6610
    0.7421    0.0043   -0.6922   -0.0350
   -0.8297   -0.0005   -0.9078    0.0007
    0.9885   -0.0269    0.0067   -0.0395
B1 =
   -0.0066
   -0.0019
    1.0183
   -0.0031
    0.5186
    0.0004
    0.3688
   -0.0004
    0.0039

W2 =
  Columns 1 through 7
    1.3803   -1.2948   -2.1872    2.6148    1.2025    0.0422    2.5109
  Columns 8 through 9
   -0.0261    3.0791
B2=
   -2.4009
```

Training set: **20 percent mortality rate (tr20)**
Test set: **true mortality rate (teorig)**

confu =

1769	90
34	361

confut =

1057	25
72	19

float = 0
layers = 2
h1 = 7
wtfact = 1

Final learning rate value
lr = 0.0023
lr_inc = 1.0030
lr_dec = 1

Final weight decay constant
lambda = 1.0000e-004
wnot = 0.1000
momentum = 0.8800
err_ratio = 1.0200
sens = 0.4318
spec = 0.9362
PPV = 0.2088
PNV = 0.9769
CCR = 0.9173

Training Set Apriori Probabilities
trprob =
79.9911
20.0089
Test Set Apriori Probabilities
teprob =
96.2489
3.7511

MDC Training Set Classification Rate
bayes_train = 83.6735
MDC Test Set Classification Rate
bayes_test = 93.0094

W1 =
Columns 1 through 7

-0.5123	-0.0109	0.4312	-0.9086	0.6742	0.9535	0.9982
-0.0131	0.0187	0.3668	-0.6644	-0.5851	0.9627	0.0160
0.5106	-0.0756	-1.9356	0.8756	0.1065	0.0355	0.0166
0.3136	-1.0002	-1.6697	-1.1612	-0.3958	0.4709	0.0351
0.8352	-0.0079	-0.0340	-0.0217	-0.3620	-0.0370	-0.0188
-0.0381	1.9433	-0.5138	-0.4894	-0.0068	0.5610	0.8218
-0.7841	0.7448	0.0069	-0.5638	0.0037	-0.0020	0.4548

Columns 8 through 14

1.0466	-0.0012	0.0011	-0.5833	0.0327	0.8237	0.9295
-1.3224	0.0233	0.0172	0.0106	1.5897	0.5382	0.0028
0.5126	0.6126	0.3063	0.7385	-1.1339	0.5960	0.0003
-1.1305	0.4981	1.2108	0.5508	-0.8105	0.8134	0.7040
-0.7093	-0.6497	0.6873	0.0359	-0.5340	-0.0196	0.0040

```
 -0.6973   0.7978  -1.3194   0.8303  -0.2703   0.0030  -0.8496
  0.6968   0.5269   0.9401  -0.7650   0.7458   0.7588   0.0031
Columns 15 through 21
  0.0011   0.0127   0.3223  -0.3740   0.2541  -0.0009   0.0107
  0.0166   0.0337   0.0250   0.5622  -0.3999  -0.0457  -1.2272
  0.0041   1.5982  -1.1928   0.0387   0.0018  -0.0888  -0.0108
  1.1214   0.1727  -0.2827  -0.0628  -0.5084   0.1483  -0.2881
 -0.0049  -1.0252  -0.0890  -0.7059   0.7785  -1.1607   0.6231
 -0.0175  -1.0842   1.0201  -0.3235   1.1881  -0.1024   0.8013
  0.0043   0.7903  -0.5804  -0.4381   0.0043  -0.0131   0.0079
Columns 22 through 28
 -0.0010  -0.3520   0.6190   0.5452   0.0011   0.8487  -0.4738
 -0.3373  -1.6025  -0.3266   1.1195   0.0267  -0.0374   0.0426
 -1.2052   0.3437  -0.2327  -0.0120   0.0035   1.1415   0.5988
  0.2141  -0.3692   0.1284  -0.2497   0.0374  -0.1992  -0.5081
  1.1968  -0.6404  -0.0115   0.1278  -0.1662   1.2877  -0.0256
  1.8349   0.8239   0.7615  -2.0441  -0.0034   0.6969  -0.5606
 -0.6978  -1.0231   0.6679   1.0401   0.9689  -0.6903  -0.0005
Columns 29 through 32
  0.7000   0.0128  -0.0011  -0.8519
  1.3965  -0.0225   0.3957   0.6999
 -0.5035  -0.7558   0.0702   0.2405
 -1.1701  -0.3988  -0.0015   1.4733
  0.4748   0.2959  -1.2048  -0.3371
  0.9485  -0.3012   0.0416   1.3551
 -0.6464   0.7160  -0.7527   0.5705
B1 =
 -0.7073
  0.6234
 -0.6141
  0.1464
  0.0052
  0.0070
  0.4243

W2 =
  1.5324   1.5313   2.2748  -3.2346  -1.1603   2.6706   0.1416
B2 =
 -0.0120
```

Training set: 10 percent mortality rate (tr10)
Test set: 20 percent mortality rate (te20)

confu =

2009	122
20	103

confut =

922	188
25	38

float = 0
layers = 2
h1 = 7
wtfact = 1.4000

Final learning rate value
lr = 1.0000e-003
lr_inc = 1
lr_dec = 1

Final weight decay constant
lambda = 4.0000e-004
wnot = 0.1000
momentum = 0.1000
err_ratio = 1.0200
sens = 0.1681
spec = 0.9736
PPV = 0.6032
PNV = 0.8306
CCR = 0.8184

Training Set Apriori Probabilities
trprob =

90.0177
9.9823

Test Set Apriori Probabilities
teprob =

80.7332
19.2668

MDC Training Set Classification Rate
bayes_train = 89.9290
MDC Test Set Classification Rate
bayes_test = 83.2055

W1 =

Columns 1 through 7

-0.6498	-0.3032	0.4604	-0.7809	0.6179	0.8549	0.7484
-0.4584	0.4815	0.6664	-0.6269	-0.7759	0.5174	0.3071
-0.3425	-0.0276	-0.8710	-0.0028	-0.7505	-0.4421	0.1348
0.4963	-0.5634	-0.7038	-1.2850	-0.6132	0.0506	0.3841
0.6468	-0.0411	-0.0208	-0.0031	-0.5860	-0.5997	0.1416
0.7509	0.6175	-0.6530	-0.4117	-0.1348	-0.0382	0.6884
-0.7478	0.5861	0.0217	-0.6365	0.0773	0.1887	0.3764

Columns 8 through 14

1.0797	-0.2668	0.1510	-0.6353	-0.0135	0.8610	0.9859
-1.0255	0.0264	-0.2859	0.0268	0.6296	0.4488	-0.0046
-0.8332	-0.0478	0.3095	0.9736	-0.4943	0.7431	-0.2217
-0.6971	-0.0364	0.6587	0.7775	-0.8225	0.2816	0.7019
-0.3359	-0.5858	0.7599	-0.0382	-1.0743	0.0044	0.0120

-0.4219 0.1656 -0.7287 0.5937 -0.3033 -0.3407 -0.6592
0.4916 0.6296 1.0553 -0.6329 0.7869 0.7696 -0.1017

Columns 15 through 21

0.2091 -0.2069 0.1671 -0.3197 0.4202 0.0362 -0.1781
-0.2016 -0.1297 0.2813 0.1582 -0.5175 0.0103 -0.8254
0.1805 0.6077 -0.3792 -0.2809 0.0058 -0.0057 0.0578
0.6257 0.0571 -0.4341 0.0702 0.7403 0.0066 0.1271
-0.0011 -0.9018 -0.0337 -0.5074 0.5567 -0.8499 0.7660
0.6954 -0.2063 0.4903 0.2397 0.3626 0.0884 0.7295
0.0071 0.5427 -0.3403 -0.4009 -0.0076 0.7530 -0.3271

Columns 22 through 28

-0.0124 -0.6427 0.6263 0.6665 -0.0179 0.8456 0.0146
-0.3787 -0.9348 0.1888 0.7616 -0.0123 0.5428 0.2817
-0.2280 0.8516 0.5868 -0.1650 0.0006 1.1895 0.6524
0.5738 -0.5547 0.5872 -0.1249 -0.8620 -0.3977 -0.1684
0.8849 -0.4703 -0.3266 0.0118 -0.5624 0.6603 0.5305
1.2752 0.0879 0.7870 -0.8136 -0.0352 0.7864 -0.2551
-0.7511 -1.2271 0.6113 1.1015 0.8351 -0.6390 -0.0137

Columns 29 through 32

0.4851 0.0496 0.2316 -0.7551
0.8943 -0.0084 0.8949 0.0996
-0.7571 -0.2558 0.6461 -0.0403
-0.0402 -0.3526 -0.0160 0.7603
0.8056 0.9054 -0.7128 -0.3779
0.2658 0.3746 -0.6959 0.8348
-0.9367 0.5274 -0.9997 0.6051

B1 =

-0.6868
0.7387
-0.5613
0.7612
-0.2659
-0.1185
0.3461

W2 =

0.9065 0.3707 -0.5194 -0.9476 -0.6071 0.9259 -0.5800

B2 =

-0.5649

Training set: 20 percent mortality rate (tr20)
Test set: 20 percent mortality rate (te20)

confu =

1525	173
278	278

confut =

795	109
152	117

float = 0
layers = 2
h1 = 2
wtfact = 1.2000

Final learning rate value
lr = 0.0013
lr_inc = 1.0030
lr_dec = 1

Final weight decay constant
lambda = 3.0000e-004
wnot = 0.1000
momentum = 0.5000
err_ratio = 1.0200
sens = 0.5177
spec = 0.8395
PPV = 0.4349
PNV = 0.8794
CCR = 0.7775

Training Set Apriori Probabilities
trprob =
79.9911
20.0089

Test Set Apriori Probabilities
teprob =
80.7332
19.2668

MDC Training Set Classification Rate
bayes_train = 83.6735

MDC Test Set Classification Rate
bayes_test = 83.3760

W1 =

Columns 1 through 7

-0.1662	-2.1414	-0.0920	0.0392	0.3195	-0.0380	-0.6661
0.0001	-0.0225	0.4616	-0.3244	0.3458	-0.0264	0.8924

Columns 8 through 14

-0.7782	-0.5474	0.1077	0.1253	-0.6043	-0.1089	0.0434
0.2195	-0.9441	-0.4033	-1.0239	-0.0523	0.0255	-0.3109

Columns 15 through 21

1.2863	-0.4660	-0.7300	-0.4354	0.6547	0.0443	-0.2784
-0.7827	0.1152	-0.8675	0.8521	-1.0054	-1.1739	-0.2341

Columns 22 through 28

0.0581	-0.1060	-0.2979	0.8219	-0.4107	-0.7110	-0.2758
0.0457	0.5596	0.8392	0.8297	-0.2232	-0.1672	0.4449

Columns 29 through 32

-0.9812	-0.0443	-0.2595	-0.1551

```
   0.0183   0.0270  -0.1776   0.1356
B1 =
  -0.8744
   0.0009

W2 =
  -1.1662  -0.5872
B2 =
  -0.4858
```

Training set: 30 percent mortality rate (tr30)
Test set: 20 percent mortality rate (te20)

confu =

1358	115
220	561

confut =

757	92
190	134

float = 0
layers = 2
h1 = 2
wtfact = 1

Final learning rate value
lr = 4.6214e-004
lr_inc = 1.0030
lr_dec = 1

Final weight decay constant
lambda = 3.0000e-004
wnot = 0.1000
momentum = 0.3500
err_ratio = 1.0200
sens = 0.5929
spec = 0.7994
PPV = 0.4136
PNV = 0.8916
CCR = 0.7596

Training Set Apriori Probabilities
trprob =
70.0089
29.9911

Test Set Apriori Probabilities
teprob =
80.7332
19.2668

MDC Training Set Classification Rate
bayes_train = 80.8784

MDC Test Set Classification Rate
bayes_test = 79.0281

W1 =

Columns 1 through 7

0.0244	-1.2732	0.4412	-0.2932	-0.6960	-0.4263	0.0130
-0.2657	1.3366	0.5147	-0.0259	-0.0171	0.0137	0.9812

Columns 8 through 14

0.9358	-0.8048	-0.0148	0.0149	-0.9365	-0.3327	-0.1491
1.3227	0.1803	0.0258	-0.1197	-0.0256	0.2611	-0.2535

Columns 15 through 21

1.0994	0.0340	-0.7008	-0.0678	0.4279	-0.6037	0.0442
-1.0557	0.1925	-0.2477	0.2230	-0.7460	-0.3814	0.5106

Columns 22 through 28

0.5767	-0.1935	0.7550	1.2471	-0.9650	0.1675	-0.1089
0.2717	0.0061	0.4542	-0.0552	0.1598	0.5195	0.1555

Columns 29 through 32

-1.6724	-0.2450	-0.5040	0.3905
0.7552	-0.0511	-0.0397	0.3366

B1 =
-0.4850
0.0177

W2 = -1.3524 1.5360
B2 = 0.3620

Appendix J: Curriculum Vitae

Candidate's Full Name	Colleen Michelle Ennett
Place and Date of Birth	Hanover, Ontario, Canada: February 11, 1974
Current Address	420 Gloucester Street, Apt. 2001, Ottawa, Ontario, Canada, K1R 7T7 Tel: 613-230-4755
Permanent Address	115 Fifth Street, Hanover, Ontario, Canada, N4N 1B5 Tel: 519-364-5390
Email Address	Ennett@canada.com

ACADEMIC QUALIFICATIONS

PhD, Systems and Computer Engineering (emphasis in Biomedical Engineering), accepted for January 2000
Carleton University,
Department of Systems and Computer Engineering,
1125 Colonel By Drive,
Ottawa, Ontario, Canada, K1S 5B6
Supervisor: Monique Frize, PhD, PEng, OC, University of Ottawa and Carleton University
Scholarships: Ontario Graduate Scholarship (OGS), Carleton University Proficiency Scholarship

MASc, Electrical Engineering (emphasis in Biomedical Engineering), 1997-99
University of Ottawa,
School of Information Technology and Engineering (SITE), Electrical Engineering
161 rue Louis Pasteur,
Ottawa, Ontario, Canada, K1N 6N5
Thesis Title: Coronary surgery mortality prediction using artificial neural networks
Supervisor: Monique Frize, PhD, PEng, OC, University of Ottawa and Carleton University

BSc (Eng), Biological Engineering, 1993-97
University of Guelph,
School of Engineering,
Guelph, Ontario, Canada, N1G 2W1
Senior thesis: Closed-loop control system for blood glucose regulation
Scholarships: Canada Scholarship for Science and Technology

John Diefenbaker Secondary School, Hanover, Ontario (1989-92)
James A. Magee Public School, Hanover, Ontario (1987-1988)
Dawnview Public School, Hanover, Ontario (1979-86)

RELEVANT WORK EXPERIENCE

Graduate Student, 1997-99

University of Ottawa

Worked as a research assistant for Dr. Monique Frize, a professor in Electrical Engineering. Also employed as a teaching assistant for all undergraduate levels of electrical engineering courses in French and English, including digital signal processing, digital electronics, and analog electronics

LIST OF PUBLICATIONS

1. Ennett CM, Frize M. An investigation into the strengths and limitations of artificial neural networks: an application to an adult ICU patient database. Proceedings of the AMIA '98 Annual Symposium, Orlando, FL, Nov 7-11, 1998:998.
2. Frize M, Wang L, Ennett CM, Nickerson B, Solven FG, Stevenson MH. New advances and validation of knowledge management tools for critical care using classifier techniques. Proceedings of the AMIA '98 Annual Symposium, Orlando, FL, Nov 7-11, 1998:553-558.
3. Tong Y, Frize M, Ennett CM. Discussion on the use of information theory to classify medical outcomes. Proceedings of the AMIA '98 Annual Symposium, Orlando, FL, Nov 7-11, 1998:1086.
4. Ennett CM, Frize M. Artificial neural network-based ICU data prediction. Ottawa Life Sciences Council Annual Conference, Ottawa, ON, Nov 17-18, 1998:abstract and poster presentation.
5. Adeney KM, Ennett CM, Frize M, Korenberg MJ. Prediction of patient outcomes in an intensive care unit. Proceedings of the IJCNN '99 Annual Symposium, Washington, DC, July 10-16, 1999.
6. Ennett CM, Frize M, Shaw RE. Methodologies for predicting coronary surgery outcomes. Proceedings of the First Joint Meeting of BMES and EMBS, Atlanta, GA, Oct 13-16, 1999.
7. Ennett CM, Frize M. (in preparation) Weight-elimination neural networks applied to coronary surgery mortality prediction. Journal of American Medical Informatics Association (JAMIA), 2000.

ELECTED POSITIONS

- Vice President for Ottawa-Carleton Chapter of WISE (Women in Science and Engineering) – professional organization: May 1999-present
- President of Electrical Engineering Graduate Students' Association at University of Ottawa: Nov 1998-Sept 1999
- Representative on the university level Graduate Students' Association (GSAÉD – Graduate Students' Association des étudiant.e.s diplômé.e.s) at University of Ottawa: Nov 1998-Sept 1999
- Program Chair for Ottawa-Carleton Chapter of WISE – professional organization: Nov 1997-May 1999
- Secretary of Engineering Society at University of Guelph: Nov 1996-Apr 1997

RESEARCH INTERESTS

- Artificial intelligence systems.
- Bioinstrumentation for measuring physiological parameters.
- Developing medical decision aid systems.
- Enhanced assistive technologies for rehabilitation engineering.
- Statistical modelling and neuro-computing.