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. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

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.

UMI®

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

Université d'Ottawa · University of Ottawa

DEVELOPING ANN APPROACHES TO ESTIMATE NEONATAL ICU OUTCOMES

Author: Yanling Tong
Master of Engineering (E.E.), Tsinghua University, 1996

Supervisor: Dr. Monique Frize, BASc(Ott), MPhil(IC), DIC (Imperial College), MBA (Moncton), PhD(Erasmus)

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Science in Engineering

Electrical and Computer Engineering, S.I.T.E.
University of Ottawa

January, 2000

ACKNOWLEDGEMENTS

Many people have supported me through times both bleak and wonderful.

First of all, I would like to give my special thanks to my supervisor, Dr. Monique Frize, for her effective guidance, kind help, great inspiration and encouragement, which is truly uncommon.

During my whole graduate program, I am grateful to my husband Dong for his constant love, patience and support.

For the development of the project itself, I feel a deep sense of gratitude to the following people:

--- Dr. Robin Walker for his valuable suggestions and timely replies.

--- my colleague Miss Colleen Ennett for her kind help, suggestions and feedback.

--- Dr. Tet Yeap for his suggestion, book, and web site on Radial Basis Function.

--- Dr. Rafik Goubran and Dr. Gibbons for their helpful ideas and feedback.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ANN	Artificial Neural Network
APACHE	Acute Physiology and Chronic Health Evaluation
BP	Back Propagation
BPD	Broncho-Pulmonary Dysphasia
CCR	Correct Classification Rate Measure of the ability of an ANN to correctly classify the test data set.
CP	Constant Predictor A statistical tool that classifies all cases as belonging to the output class with the highest probability.
CRIB	Clinical Risk Index for Babies
EM	Expectation-Maximization
ICD9	Classification of Diseases 9
ICU	Intensive Care Unit
IVH	Intra-Ventricular Hemorrhage
LOS	Length Of Stay
NICU	Neonatal Intensive Care Unit
NN	Neural Network
NP	Nondeterministic Polynomial time. Definition NP is the class of decision problems (languages) L such that there is a polynomial time function $f(x,c)$ where x is a string, c is another string whose size is polynomial in the size of x, and $f(x,c) = True$ if and only if x is in L.
NP-hard	A problem that is at least as hard as or harder than any problem in NP, where NP means nondeterministic polynomial time
NTISS	Neonatal Therapeutic Intervention Scoring System
ROC	Receiver Operator Curves
ROP	Retinopathy Of Prematurity
SAPS	Simplified Acute Physiology Score
SNAP	Score for Neonatal Acute Physiology
TISS	Therapeutic Intervention Scoring System

DEFINITIONS

Chi-square Test for Goodness of Fit

The chi-square test for goodness of fit tests the hypothesis that the distribution of the population from which nominal data are drawn agrees with a posited distribution. The chi-square goodness-of-fit test compares observed and expected frequencies (counts). The chi-square test statistic is basically the sum of the squares of the differences between the observed and expected frequencies, with each squared difference divided by the corresponding expected frequency.

Chi-square Test for Independence (Pearson's)

Pearson's chi-square test for independence for a contingency table tests the null hypothesis that the row classification factor and the column classification factor are independent. Like the chi-square goodness-of-fit test, the chi-square test for independence compares observed and expected frequencies (counts). The expected frequencies are calculated by assuming the null hypothesis is true. The chi-square test statistic is basically the sum of the squares of the differences between the observed and expected frequencies, with each squared difference divided by the corresponding expected frequency. Note that the chi-square statistic is always calculated using the counted frequencies. It can not be calculated using the observed proportions, unless the total number of subjects (and thus the frequencies) is also known.

Correlation Correlation is the linear association between two random variables X and Y. It is usually measured by a correlation coefficient, such as Pearson's r, such that the value of the coefficient ranges from -1 to 1. A positive value of r means that the association is positive; i.e., that if X increases, the value of Y tends to increase linearly, and if X decreases, the value of Y tends to decrease linearly. A negative value of r means that the association is negative; i.e., that if X increases, the value of Y tends to decrease linearly, and if X decreases, the value of Y tends to increase linearly. The larger r is in absolute value, the stronger the linear association between X and Y. If r is 0, X and Y are said to be uncorrelated, with no linear association between X and Y. Independent variables are always uncorrelated, but uncorrelated variables need not be independent.

Goodness of Fit

Goodness-of-fit tests test the conformity of the observed data's empirical distribution function with a posited theoretical distribution function. The chi-square goodness-of-fit test does this by comparing observed and expected frequency counts. The Kolmogorov-Smirnov test does this by calculating the maximum vertical distance between the empirical and posited distribution functions.

ABSTRACT

A medical database is a valuable resource for medical research. By analyzing the patient records, the physicians can know better about the patient outcomes and resources utilization. In NICU where the medical resources are very expensive, it is particularly important for the physicians to know the relationship between the patient measurements and the outcomes (e.g. death, ventilation, length of stay, and other complications including lung disease and brain damage).

Many techniques have been used in predicting or estimating patient outcomes in the literature. In Chapter 2 these approaches were compared in performance and complexity. A backpropagation neural network (NN) was selected. Because neural networks cannot handle missing information, whereas incomplete patient records is a very common occurrence in the Neonatal Intensive Care Unit (NICU) database, two sets of experiments were conducted.

In the first set of experiments, all the cases with missing information in a NICU database of about seven thousand patients were excluded. Only complete records were kept. Encouraging results were obtained using neural networks in predicting ventilation (<=8hours or >8hours) with SNAP II scoring system variables. The effectiveness of weight-elimination in controlling over-fitting was also validated. By comparing the result from an adult ICU (AICU) database, it was verified that the adult model could be successfully applied to neonates.

In the second set of experiments, we tended to make full use of the records with missing information. The missing values were replaced with their NORMAL values. Unfortunately the cases with lower apriori probability were classified poorly by NN. But to solve this challenging problem, much effort and many other methods are needed. For example, the large difference of the mean of some variables in the databases with opposite outputs identified by statistical comparison implied that there should be a way to separate these classes.

Finally, a backpropagation neural network model was implemented with C++. The network parameters can be adjusted outside the source file, which facilitates the experiments and decreases the training period compared to the approach using MATLAB developed by Trigg [Trigg, 1997]. The program has been tested with one adult ICU database.

At the end, some future research is proposed.

1. INTRODUCTION

Currently in medicine, health care organizations are increasingly faced with demands for information about the quality and cost of health care services. Various scoring systems have been developed to assess patient outcomes in critical care; examples are APACHE, SAPS, TISS, NTISS, CRIB, and SNAP.

The "Acute Physiology And Chronic Health Evaluation" (APACHE) severity of disease classification system (Knaus *et al.*, 1981) and later "APACHE II" are among the first. The "Simplified Acute Physiology Score" (SAPS) and its follow-up version SAPS II were designed (LeGall *et al.*, 1995) to improve the predictability of hospital mortality. Another system, "Therapeutic Intervention Scoring System" (TISS) has been used in the ICU (Intensive Care Unit) to analyze expenditures, outcomes and care. In the NICU (Neonatal Intensive Care Unit), several scoring systems have also been adopted, namely, "Neonatal Therapeutic Intervention Scoring System" (NTISS), "Clinical Risk Index for Babies" (CRIB), and "Score for Neonatal Acute Physiology" (SNAP).

Richardson and Tarnow-Mordi (1994) reviewed 30 neonatal scoring systems and found that few had been validated on large, concurrent samples of newborns. Knowledge-based systems are being increasingly accepted as part of clinical decision-support systems (Yu *et al.*, 1982; Dawant *et al.*, 1993; Lau, 1994); Johnston *et al.*, in 1994, concluded that clinical decision-

support systems could improve clinician performance and patient outcomes in clinical settings such as quality assurance for active medical care. They also pointed out the need for new, well-designed studies to assess the effects and the cost-effectiveness of clinical decision-support systems, especially when attempting to affect patient outcomes [Frize *et al.*, 1995].

1.1 Medical Background

Collecting medical information is the first important step in health care research. Between January 8, 1996 to October 31, 1997, there were about 20,000 admissions of babies to 17 Canadian Neonatal Intensive Care Units (NICUs). This did not include the babies who were transferred to a normal newborn care area within 24 hours [Lee, 1999].

Once the patient was admitted to the NICU, the doctor prescribed a set of tests or treatment according to the basic measurements (such as blood pressure). Most patients recovered after being given appropriate treatment, and many can be treated effectively with less time in the NICU. It is very important to make a medical decision accurately, swiftly, and economically, and identify patients at high risk from diseases like intra-ventricular hemorrhage. The patients at high risk may be hospitalized to receive aggressive testing and treatment, while patients at low risk may be more comfortably, safely, and economically treated in a normal baby care area or at home [Caruana, 1996].

This requirement has led to the interest in determining the relative importance of various factors contributing to mortality and long-term morbidity of babies. In neonatal intensive care, it is required to develop a simple and reliable score of illness severity for neonates. Decision-

support systems are desired to assist clinicians in making difficult treatment decisions by accessing the cumulative experience of clinicians from all over the country [Frize, 1998].

1.2 Three Main Neonatal Scoring Systems

The most recognized scoring systems currently used in NICUs are: Neonatal Therapeutic Intervention Scoring System (NTISS), Score for Neonatal Acute Physiology (SNAP), and Clinical Risk Index for Babies (CRIB).

NTISS is a neonatal variation of the TISS system. The admission day NTISS scores are found to correlate with mortality, Length Of Stay (LOS), and total hospital charges for survivors in the USA [Gray *et al.*, 1992].

CRIB and SNAP are more popular than NTISS in neonatal research. They use demographic and physiologic variables rather than therapeutic interventions. CRIB is based on routine data recorded within 12 hours of birth. It was validated only in a cohort of infants with birth weight 1500 gram or less, or gestational age less than 31 weeks [INN, 1993]. CRIB correlates with morbidity and risk of mortality, but less well with LOS [DeCourcy *et al.*, 1995]. SNAP is a scoring system modeled on APACHE. It was developed to facilitate risk-adjusted comparisons of mortality between NICUs. Since SNAP contains more variables than CRIB, it may prove more valuable for research than for routine use [INN, 1993]. It is reported to correlate well with mortality and LOS in NICUs, particularly for infants under 2500 gram [Escobar *et al.*, 1995].

These scoring systems are static systems, which means that they use the patient admission information only. Whether the temporal characteristics could add relevance and usefulness for medical practitioners has not yet been investigated.

1.3 Previous Work by Medical Ideas Research Group (MIRG)

The Medical IDEAS Research Group (MIRG) is a multidisciplinary research group whose members are investigating a number of promising information technology approaches to assist with decision-making and patient management in the medical environment [Frize *et al.*, 1995, 1996]. One of the main projects is to employ ANNs to estimate medical outcomes and resource utilization. This research is a part of it.

MIRG's preliminary work in using ANNs yielded encouraging results [Buskard *et al.*, 1994]. Buskard used a backpropagation ANN to estimate "Mortality", "Length of Stay" and "Ventilation" for patients in a database of 1322 Medical/Surgical Intensive Care Unit patients collected at the Doctor Everett Chalmers Hospital (DECH). He obtained slightly better estimates of correct classification rates (CCRs) (89.4%, 66.1% and 69%, in "Mortality", "Length of Stay" and "Ventilation", respectively) than a Constant Predictor did (CCRs of 88.8%, 61% and 66.5% respectively). Constant Predictor (CP) is a statistical tool that classifies all cases as belonging to the output class with the highest probability. But he observed that after a few hundred training epochs, the classification rate and error curves for the training and test sets diverged and the double-layered networks did not improve upon the performance levels observed for single-layered network. This is called overfitting.

This overfitting problem was successfully solved by another graduate student - Trigg [Trigg, 1997] at the University of New Brunswick. A technique named 'weight elimination', which adds a penalty term to the standard backpropagation cost function to force already-small weights to zero, was used to improve the generalization ability of the ANNs. By this method, the correct classification rates of ANNs for the outcome 'duration of ventilation' less than or equal to or more than 8 hours for postoperative ICU patients were improved by 18.8 to 20.7% compared to a Constant Predictor (CP). The best ANN classification performance was also 2.4 to 5.7% better than classification using a Minimum-Distance Classifier.

Trigg's work was further improved by Ennett [1998] and Ho [1998]. Six variables were extracted from 51 variables by using a one-layered ANN with weight elimination technique. This simplified network generally presented a better performance in terms of CCR, ASE and training time.

2. LITERATURE REVIEW

A number of different methods have been used to model medical data. Classification analysis is equated with supervised learning; clustering is equated with unsupervised learning; and regression is equated with functional analysis. They range from the non-flexible (logistic regression), through partially flexible (Generalized Additive Models or GAMs), to completely flexible (classification trees and neural networks) [Plate *et al.*, 1997]. Before proceeding with any of these methodologies, data preprocessing techniques should be used to understand the data distribution and clean the database.

2.1 Data Preprocessing

"Neural networks are a good way to interrelate nonlinear variables in a robust manner. But if the training data set (the collection of input data and its associated correct output data) is not thoughtfully chosen, the resulting network is unlikely to hold up well in an industrial environment. It is hardly surprising that massaging the set of training data consumes some 80 percent of the engineering time spent getting a real-world neural network up and running – that is, getting it to converge under a broad enough range of conditions to be deployed with confidence in a production situation. If that data preparation is done systematically, much time can be saved and a more end-product can be obtained." These words published on "IEEE

Spectrum" are from the experience of Yale [1997]. They pointed out the importance of carrying on data preprocessing systematically. Actually, the old adage "garbage in ... garbage out" applies as much to all other data-processing applications as it does to neural networks. Therefore, data preprocessing should be carried on very carefully.

SPSS is a comprehensive statistical software system designed to handle all steps in an analysis ranging from data listings, tabulations, and descriptive statistics to complex statistical analyses. With its help, data preprocessing can be done in a step by step manner.

2.1.1 Analysis of Individual Variables

- **Frequency Analysis**

Frequency analysis is usually the first step to study the database. By performing a univariate analysis, this procedure analyzes only one variable at a time. The calculation for each variable includes counts, percentages, mean, median, standard deviation, max, min, histogram, *etc.*. From the values and histograms, it is easy to view the distribution of the variables. It is also helpful to find out the missing and extreme values.

In order to simplify the database, some records and variables can be removed, such as records that do not have many meaningful observations for any variable, or variables that have too few observed data associated with it.

- **Declare Missing Values**

Missing values are values or code numbers that represent missing information. There are two kinds of missing values: system missing values and user-defined missing values. A system missing value is created when SPSS cannot find or interpret a variable's value for one or more cases in the data file. A user-defined variable is coded by the user as a special symbol (e.g., '-9') which means that the variable is missing or left null. Since it is meaningless to incorporate them to calculate statistics such as correlation coefficients, they can be simply treated as missing values in data analysis. These data records cannot be totally eliminated because, on one hand, the total amount of remaining data may not be sufficient and, on the other hand, the remaining values in the data record may contain very useful information. But, traditionally, if more than 20% of attribute values are missing, the entire record should be eliminated [Famili *et al*, 1997].

- **Find Outliers**

An outlier or "outside values" among a set of residuals is much larger than the rest in absolute value, perhaps lying three or more standard deviations from the mean of the residuals. It seems quite out of place with (considerably higher or lower than) the other data. Clearly, the presence of such an extreme value can significantly affect the least-squares fitting of a model, so it is important to determine whether the outlier should be removed from the data [Kleinbaum, Kupper, Muller, *et al.* 1998]. The following approaches can be used to detect an outlier:

1) Box plot and stem-and-leaf diagram;

2) Scatter plot (from the scatter plot, the necessary data transformation techniques can be decided, e.g., log transformation or square transformation, *etc.*);

3) Diagnostic regression statistics for evaluating outliers: jackknife residuals, leverages, and Cook's distance (a measure of influence).

2.1.2 Correlation

Generally, the variables in a database can be classified into two categories: quantitative variables and qualitative variables. Quantitative variables (such as "high blood pressure" and "respiration rate") describe the attributes in terms of numerical measurement. Qualitative variables describe the attributes that are inherently non-numeric or that have been measured according to standards that do not lend themselves to numerical expression. Qualitative variables can be divided into categorical or nominal variables and ordinal variables.

The Pearson correlation method can be used to test the relationship among the quantitative variables that are normally distributed. The Spearman correlation method can be used for ordinal variables and quantitative variables that are not guaranteed to have a normal distribution. Chi-square procedures can be used for nominal variables.

2.1.3 Principal Components Analysis

In complex data analysis applications, although there may be hundreds of measurements, relatively few events may be occurring. Fayyad *et al.* [1996] emphasized the use of data preprocessing techniques as an essential part of any knowledge discovery from a data base

project. For example, in classification with neural networks, one can eliminate irrelevant data to produce faster learning due to smaller data sets and reduction of confusion caused by irrelevant data.

Complexity may be significantly reduced if irrelevant data are eliminated and only the most relevant features are used for data analysis (e.g., by principal components analysis) to narrow the search space. Reducing the dimensionality (through eliminating irrelevant data) may also improve the performance of a data analysis tool, since the number of training examples needed to achieve a desired error rate increases with the number of input variables [Famili *et al.* 1997]. Linear transformations such as the principal component transformation may prove useful for this.

The use of principal components has been extensively studied in [Duszak and Loczkodaj, 1994]. The main goal of identifying principal components is to select proper attributes for data analysis. Theoretically, selecting X attributes from Y ($X<Y$) is equivalent to selecting X basis vectors, spanning the subspace on these X vectors, and projecting the database onto this space. Therefore, identifying principal components allows to reduce the dimensionality of a database in which there are large numbers of interrelated variables, while retaining as much as possible of the variables present in the database. This reduction is achieved by transforming to a new set of variables, called principal components, which are highly uncorrelated, and which are ordered so that the first few retain most of the variations present in all of the original variables. Identifying principal components involves checking the linear dependency among independent variables in a

set of data attributes, which can be done automatically as part of the data analysis or separate from the analysis process [Famili *et al.*, 1996].

In brief, the principal component analysis (PCA) may achieve three concomitant goals [Frize, 1998]. First it can transform the data from dimension Y to dimension X (<Y), without much loss of information, i.e., the variance structure can be approximately reconstructed with fewer variables. Second the interpretations of the coefficients and subsequent viewing of the principal component images can aid in discerning important features. Third, the resulting variables are uncorrelated.

2.2 Logistic Regression

So far the patient outcomes predicted in this research are in categorical forms (e.g. mortality=0 means survival, mortality=1 means non-survival). Therefore, the techniques for analyzing data with categorical dependent variables (e.g. ***discriminant analysis***, ***probit analysis***, ***log-linear regression*** and ***logistic regression***) are of interest.

These various techniques are applicable in different situations: for example, log-linear regression requires all regressors to be categorical, whilst discriminant analysis strictly requires them all to be continuous (though dummy variables can be used for multiple regression). Logistic regression is used when the dependent is a dichotomy and the independent variables are continuous, categorical, or both. Discriminant analysis is fairly often encountered in journals. But it is now being replaced with logistic regression, as this approach requires fewer assumptions in theory, is more statistically robust in practice, and is easier to use and understand than

discriminant analysis [Press and Wilson, 1978]. In SPSS at least, logistic regression is easier to use than discriminant analysis when we have a mixture of numerical and categorical regressors, because it includes procedures for generating the necessary dummy variables automatically.

Logistic regression does not assume linearity of relationship between the independent variables and the dependent, does not require normally distributed variables, does not assume homoscedasticity, and in general has less stringent requirements. It is the most widely used approach in medical research. Looking at the classification table, and, showing correct/incorrect classifications of the dichotomous dependent can assess the success of the logistic regression. Also, goodness-of-fit tests are available as indicators of success.

In logistic regression, the effects of predictors are linear (or categorical) and additive on the log-odds scale. The logistic regression model for a binary dependent variable is:

$$E(y) = \frac{e^{b_0 + b_1 + b_2 + ... + b_k}}{1 + e^{b_0 + b_1 + b_2 + ... + b_k}}$$, where coefficients $b_1, b_2, ..., b_k$ measure the regressor's independent contribution to variations in the dependent variable $E(y)$. Instead of using a least-squared deviations criterion for the best fit as linear regression does, logistic regression finds a "best fitting" equation with maximum likelihood method, which maximizes the probability of getting the observed results given the fitted regression coefficients.

An important virtue of logistic regression is that the relationships identified in the data can be interpreted and explained in simple terms. For example, the result can be explained as "The odds of developing lung cancer for males who smoke between 20 and 29 cigarettes per day increased by a factor of 11.5 over males who do not smoke." [Plate *et al.*, 1997].

The odds ratio is a widely used measure of effect in epidemiological studies. By "measure of effect," we mean a measure that compares two or more groups in predicting the outcome (dependent) variable. To describe an odds ratio, we first define odds as the ratio of the probability that some event (e.g., developing lung cancer) will occur divided by the probability that the same event will not occur (e.g., not developing lung cancer). Thus, the odds for some event D is given by the formula [Kleinbaum, Kupper, Muller, *et al.* 1998]:

$$odds(D) = \frac{pr(D)}{pr(notD)} = \frac{pr(D)}{1 - pr(D)}$$

2.3 Expert Systems

The expert systems are based on the knowledge of the experts. Several authors used the hierarchical classification system to perform medical diagnosis when building medical expert systems, or rule-based systems [Weiss *et al.*, 1978]. Although performance may be acceptable, problems with expert systems usually occur during the knowledge-acquisition phase, when a great amount of time is spent on extracting information from the expert [Forsythe and Buchanan, 1989]. Furthermore, expert judgement may contain biases [Tversky and Kahneman, 1974], a problem that machine-learning approaches may avoid by extracting information from evidence. The other problem is the inconsistencies that may arise when new rules are added to an existing database. There is a strong domain dependence; knowledge bases can rarely be reused for other applications.

2.4 Artificial Neural Networks (ANNs)

"As many IEEE Spectrum readers are doubtless aware, neural networks are a good way to interrelate nonlinear variables in a robust manner. The reason, in a nutshell, is that neural networks are put through a training phase, during which they can automatically fine-tune themselves as often as proves necessary to get the desired performance." [Yale, 1997].

Since Artificial Neural Networks (ANNs) were introduced into the commercial market, the ANN software packages have been widely used by many researchers and engineers from all disciplines because of their powerful non-linear modeling characteristics. In the medical area, the advantages of selecting ANNs over traditional statistical analyses for vast and complex databases of medical information has been recognized in several publications. In the early 1990's, ANNs were used to identify problems such as heart attacks and microcalcifications on mammographic x-rays, with varying degrees of success [Baxt, 1990 and 1991; Wu, 1993]. Baxt used it as an aid in the diagnosis of acute coronary occlusion [1990] and myocardial infarction [1991]. Later ANNs were used to diagnose other diseases. Alonzo-Betanzos (1992) focused on prenatal outcomes. Kuntz (1994) designed a cascade-correlation ANN to estimate mortality and length of stay for patients with closed-head injuries. Buchman [1994] estimated chronicity in a surgical intensive care unit. Tu & Guerriere [1993], Buskard et al. [1994] and Frize et al. [1995] reported the estimations of LOS and mortality. Buskard et al. [1994], Frize et al. [1995, 1996 and 1997] and Trigg [1997] reported the studies of estimated duration of artificial ventilation. The use of mixed models for clustered binary outcome data applied to hospital outcomes research and the use of non-linear mixed models for continuous measurements was examined in [Brandt, 1997].

Later ANNs were used to diagnose other diseases, such as, psychiatric disease [Lowell, 1998], ectopic pregnancies [McAfee, 1998], skin cancer [Lee, 1996]. Lee's system had a diagnostic accuracy of 93% while general dermatologists had an average diagnosis accuracy of 82%, benign tumors and malignant melanomas [Ma *et al.*, 1998], acute appendicitis [Pesonen, 1996, 1997], labeled monodispersed aerosols in obstructive lung disease [Delogu, in press], acute abdominal pain [Pesonen, 1998], *etc.*

2.4.1 Backpropagation Neural Network

Backpropagation (BP) neural networks were created by generalizing the Widrow-Hoff learning rule to multiple-layer networks and nonlinear differentiable transfer functions [MATLAB, 1994]. They are the benchmark against which other neural network architectures are conventionally compared. Architecture variables include the number of neuronodes and number of associative layers. These back propagation neural networks modify their connection weights through an iterative process. Trained BP neural networks tend to give reasonable answers when presented with inputs that they have never seen. Typically, a new input will lead to an output similar to the correct output for input vectors used in training that are similar to the new input being presented.

However the BP training may lead to a local rather than a global error minimum. This might be avoided by changing the initial conditions or the number of neurons and layers. Moreover, caution must be exercised in selecting these structural features, as well as the duration

of training, because back propagation neural networks are predisposed to memorize (as opposed to generalize from) training sets [Buchman *et al.*, 1994]. This phenomenon is called over-fitting.

2.4.2 Probabilistic Neural Networks

Probabilistic neural network architectures are fundamentally different from backpropagation neural networks in that the formers are "single training-pass" estimators of probability density functions. They produce probabilities as the primary outputs. In particular, there is no iteration and associated risk of memorization: a combination of Bayes strategies and Parzen windows is used. Generalization is optimized by adjusting a "smoothing factor", δ, which controls the degree of interpolation between input variable patterns presented to the network. Such networks may be trained extremely quickly. However, the entire training set must be stored and used during testing; the amount of computation necessary to classify an unknown point is proportional to the size of the training set. As such, probabilistic neural networks do not yet represent an overall speed advantage over standard backpropagation methods [Buchman *et al.*, 1994].

2.4.3 Bayesian Inferred Neural Network (NN) Classifier

According to [Roberts *et al.* 1994], outlier detection is important to get reliable results in a critical (e.g. medical) environment. To get reliable results, one must not classify the dubious inputs. In order to detect such dubious inputs, Sykacek, Dorffner *et al.* [1998] used Bayesian inference to calculate a distribution over the NN weights in classifying four sleep stages.

The Bayesian solution for NNs Sykacek, Dorffner, *et al.* [1998] used is *a posteriori* distribution over weight space calculated via Bayes' theorem using *a priori* over weights.

$$P(w \mid D) = \frac{P(D \mid w)p(w)}{P(D)}$$

where w is the weight of the network and D represents the training data. In order to calculate the a *posteriori*, a hybrid Monte Carlo method [Neal, 1996] was used to sample from the *a posteriori*.

The classifier that they used in conjunction with Bayesian inference was a 2-layer neural network with 10 inputs, 25 hidden units with sigmoid activation and 5 output units with softmax activation [Bishop, 1995]. Its mean classification performance under different percentage of dubious cases (0, 5%, 10%, and 15%) was 78.4% to 83.6%. A comparison with results achieved with neural networks optimized with steepest descent and early stopping showed a range of 50% to 65%.

They demonstrated that using more sophisticated training methods like Bayesian Inference in neural networks leads to better classification results compared with simpler training procedures. The main problem of Bayesian techniques is the large amount of time that is required to calculate the solution. From a practical point of view, a faster training method is desirable.

2.4.4 Dealing with an Incomplete Database

Most feed-forward ANNs estimate *a posteriori* probabilities with a suitable encoding of the output units. They can classify very well, but they need full information about a case. They cannot handle missing data, neither during learning nor during the classification. Success or failure often depends on the input representation.

Classification with unknown inputs has recently been addressed in the ANN context. Mixed models of normal distributions were used whereby simple closed form solutions to optimal regression with missing data can be formulated. The Expectation-Maximization algorithm (EM) [Dempster *et al.*, 1977] for parameter estimation is especially interesting in this context since it can also be formulated to handle missing data in the training examples.

2.4.5 Improving the Backpropagation Algorithm

An important focus of neural network research is the question of how to adjust the weights of the links to get the desired system behavior. This modification is very often based on the Hebbian rule, which states that a link between two units is strengthened if both units are active at the same time. The standard backpropagation learning algorithm introduced by [Rumelhart and McClelland, 1986] is the most common learning algorithm.

Since the backpropagation learning algorithm [Rumelhart, Hinton and Williams, 1986A; 1986B] was first popularized, there has been considerable research on methods to accelerate the convergence of the algorithm. This research falls roughly into two categories. The first category

involves the development of ad hoc techniques (e.g., Vogl *et. al.*, 1998; Jacobs, 1998). These techniques include such ideas as varying the learning rate, using momentum and rescaling variables. Another category of research has focused on standard numerical optimization techniques (e.g., Shanno, 1990; Barnard, 1992; Battiti, 1992; Charalambous, 1992).

Momentum decreases backpropagation's sensitivity to small details in the error surface. This helps the network avoid getting stuck in shallow minima, which would prevent the network from finding a lower error solution.

Training time can also be decreased by the use of an adaptive learning rate which attempts to keep the learning step size as large as possible while keeping the learning rate stable. The learning rate is made responsive to the complexity of the local error surface.

The most popular approaches from the second category have used conjugate gradient or quasi-Newton (secant) methods. The quasi-Newton methods are considered to be more efficient, but their storage and computational requirements go up as the square of the size of the network. There have been some limited memory quasi-Newton (one step secant) algorithms that speed up convergence while limiting memory requirements [Battiti, 1992; Liu and Nocedal, 1989]. If exact line searches are used, the one step secant methods produce conjugate directions.

Another area of numerical optimization that has been applied to neural networks is nonlinear least squares [Kollias and Anastassiou, 1989; Singhal and Wu, 1989; Puskorius and Feldkamp, 1991]. The more general optimization methods were designed to work effectively on all sufficiently smooth objective functions. However, when the form of the objective function is known it is often possible to design more efficient algorithms. One particular form of objective

function that is of interest for neural networks is a sum of squares of other nonlinear functions. The minimization of an objective function of this type is called nonlinear least squares.

While backpropagation is a steepest descent algorithm, the Levenberg-Marquardt algorithm [Marquardt, 1963] is an approximation to Newton's method. Levenberg-Marquardt optimization is a more sophisticated method than gradient descent (backpropagation). For very large networks, the memory requirements of this algorithm make it impractical for some machines (as is the case for the quasi-Newton methods). However, for networks with a few hundred weights, the algorithm is very efficient.

"Backpropagation is used in perhaps 80% to 90% of practical applications. Improvement techniques can be used to make backpropagation more reliable and faster by a factor of at least ten to twenty on small problems and perhaps more than that on larger problems." [MATLAB Neural Network Toolbox]. Some of the most popular algorithms and their main differences are discussed in the following sections.

2.4.4.1 Standard Backpropagation (Generalized Delta-rule)

Equation: $\Delta w_{ij} = \eta * \delta_j * o_i$, where

$$\delta_j = \begin{cases} f_j'(net_j)(y_j - o_j), & if \quad unit \quad j \quad is \quad a \quad output-unit \\ f_j'(net_j)\sum_k \delta_k w_{jk}, & if \quad unit \quad j \quad is \quad a \quad hidden-unit \end{cases}$$

η learning rate (a constant)

δ_j error (difference between the real output and the desired output) of unit j.

y_j desired output of unit j.

o_i output of the preceding unit i.

i index of a predecessor to the current unit j with link w_{ij} from i to j.

j index of the current unit.

k index of a successor to the current unit j with link w_{jk} from j to k.

1. Apply the input vector $x_p = (x_{p1}, x_{p2}, ..., x_{pN})^t$ to the input units.
2. Calculate the net-input values to the hidden layer units: $net_{pj}^h = \sum_{i=1}^{N} w_{ji}^h x_{pi} + \theta_j^h$
3. Calculate the outputs from the hidden layer: $i_{pj} = f_j^h(net_{pj}^h)$
4. Move to the output layer. Calculate the net-input values to each units: $net_{pk}^o = \sum_{j=1}^{L} w_{kj}^o x_{pj} + \theta_k^o$
5. Calculate the outputs: $o_{pk} = f_k^o(net_{pk}^o)$
6. Calculate the error terms for the output units: $\delta_{pk}^o = (y_{pk} - o_{pk}) \cdot f_k^{o'}(net_{pk}^o)$
7. Calculate the error terms for the hidden units: $\delta_{pj}^h = f_j^{h'}(net_{pj}^h) \sum_k \delta_{pk}^o w_{kj}^o$

   N.B. the error terms on the hidden units are calculated before the connection weights to the output-layer units have been updated.

8. Update weights on the output layer: $w_{kj}^o(t+1) = w_{kj}^o(t) + \eta \delta_{pk}^o i_{pj}$
9. Update weights on the hidden layer: $w_{ji}^h(t+1) = w_{ji}^h(t) + \eta \delta_{pj}^h x_i$

The error term $E_p = \frac{1}{2}\sum_{k=1}^{M} \delta_{pk}^2$ is calculated to measure how well the network is learning.

2.4.4.2 Enhanced Backpropagation with Momentum

This algorithm uses a momentum term and flat spot elimination. The momentum term α introduces the old weight change as a parameter for the computation of the new weight change. This avoids oscillation problems common with the regular backpropagation algorithm when the error surface has a very narrow minimum area. The effect of these enhancements is that flat spots of the error surface are traversed relatively rapidly with a few big steps, while the step size is decreased as the surface gets rougher. This adaptation of the step size increases learning speed significantly.

Equation: $\Delta w_{ij}(t+1) = \eta * \delta_j * o_i + \alpha \Delta w_{ij}(t)$

2.4.4.3 Enhanced Backpropagtion with weight-decay and weight-elimination

Weight-decay was introduced by P. Werbos [1988]. It decreases the weights of the links while training them with backpropagation. In addition to each update of a weight by backpropagation, the weight is decreased by a part d of its old value. The effect is similar to the pruning algorithms. Weights are driven to zero unless reinforced by backpropagation.

Equation: $\Delta w_{ij}(t+1) = \eta * \delta_j * o_i - d w_{ij}(t)$

Weight-decay is a special case of weight-elimination, which employs a different penalty term from weight-decay. According to Weigend *et al.* [Weigend *et al.*, 1991], the learning rule of

weight-elimination is to change the weights according to the gradient of the entire cost function. The main difference between weight-decay and weight-elimination is that "decay using the sum of squared weights tends to shrink the large coefficients more than the small ones, [and] weight-elimination tends to shrink the small coefficients more" [Sarle, 1996]. Therefore, weight-elimination is more useful for network pruning.

2.4.4.4 Levenberg-Marquart Algorithm

Training a neural network is, in most cases, an exercise in numerical optimization of usually nonlinear objective function ("objective function" is a function to be optimized and it is a slightly more general term than "error function" in that it may include other quantities such as penalties for weight decay). For objective functions with continuous second derivatives (such as feedforward NNs with the most popular differentiable activation functions and error functions), three general types of algorithms have been found to be effective for most practical purposes:

- For a small number of weights, stabilized Newton and Gauss-Newton algorithms, including various Levenberg-Marquardt and trust-region algorithms, are efficient.
- For a moderate number of weights, various quasi-Newton algorithms are efficient.
- For a large number of weights, various conjugate-gradient algorithms are efficient.

Suppose that we have a function $V(x)$ which we want to minimize with respect to the parameter vector x, then Newton's method would be: $\Delta x = -[\Delta^2 V(x)]^{-1} \nabla V(x)$, where $\Delta^2 V(x)$

is the Hessian matrix and $\nabla V(x)$ is the gradient. If we assume that $V(x)$ is a sum of squares function $V(x)=\sum_{i=1}^{N} e_i^2(x)$, then it can be shown that

$$\nabla V(x) = J^T(x)e(x)$$
$$\nabla^2 V(x) = J^T(x)J(x) + S(x)$$

where $J(x)$ is the Jocobian matrix

$$J(x)=\begin{bmatrix} \frac{\partial e_1(x)}{\partial x_1} & \frac{\partial e_1(x)}{\partial x_2} & \cdots & \frac{\partial e_1(x)}{\partial x_n} \\ \frac{\partial e_1(x)}{\partial x_1} & \frac{\partial e_2(x)}{\partial x_2} & \cdots & \frac{\partial e_2(x)}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial e_{N1}(x)}{\partial x_1} & \frac{\partial e_N(x)}{\partial x_2} & \cdots & \frac{\partial e_N(x)}{\partial x_n} \end{bmatrix}$$

and $S(x)=\sum_{i=1}^{N} e_i(x)\nabla^2 e_i(x)$. For the Gauss-Newton method, it is assumed that $S(x)\approx 0$, and then $\Delta x=-[J^T(x)J(x)]^{-1}J^T(x)e(x)$. The Levenberg-Marquardt modification to the Gauss-Newton method is: $\Delta x=-[J^T(x)J(x)+\mu I]^{-1}J^T(x)e(x)$. The parameter μ is multiplied by some factor (β) whenever a step would result in an increased $V(x)$. When a step reduces $V(x)$, μ is divided by β. Notice that when μ is large the algorithm becomes a steepest descent (with step $1/\mu$), while for small μ the algorithm becomes a Gauss-Newton.

The Levenberg-Marquardt modification to the backpropagation algorithm by Hagan and Menhaj [1994] thus proceeds as follows:

1. Present all inputs to the network and compute the corresponding network output and errors. Compute the sum of squares of errors over all inputs.

2. Compute the Jocobian matrix.
3. Solve $\Delta x = -[J^T(x)J(x) + \mu I]^{-1} J^T(x)e(x)$ to obtain Δx by Cholesky factorization.
4. Recompute the sum of squares of errors using $x + \Delta x$. If this new sum of squares is smaller than that computed in step 1, then reduce μ by β, let $x = x + \Delta x$, and go back to step1. If the sum of squares is not reduced, then increase μ by β and go back to step 3.
5. The algorithm is assumed to have converged when the norm of the gradient is less than some predetermined value, or when the sum of squares has been reduced to some error goal.

If sigmoid is used as the transfer function, then the latter can be represented as $f(x) = \frac{1}{1+e^{-x}}$ and its first derivative is $f'(x) = f(x) * (1 - f(x))$ [Rogers, 1997].

If tansig is used as the transfer function, then it can be represented as $f(x) = \frac{2}{1+e^{-2x}} - 1$ and its first derivative is $f'(x) = 1 - f(x) * f(x)$.

2.4.6 Two Key Problems in ANNs

One of the problems with ANNs is overfitting, resulting in excellent fitting to existing data but not necessarily good predictive performance for new cases. Several approaches have been used to control the overfitting: the weight elimination method applied by Trigg's [1997], Breiman's [1996] bagging technique, early stopping method and Levenberg-Marquardt algorithm.

The other difficulty in analyzing medical information is the missing data. There are several ways to solve this. The first one is that cases with missing data are not included so that

the prediction models can be compared without putting any prediction model at a disadvantage [Burke, 1996]. The second one is to replace the missing data with their mean values. But it has been reported that the substitution of the mean for an unknown feature can lead to solutions that are far from optimal [Ahmad and Tresp, 1993; Tresp, Ahmad and Neuneier, 1994]. The third one is to replace the missing data with their "normal" values as Trigg [1997] did. The last one was suggested by Mackay [1995] as a Bayesian neural network methodology. Plate *et al.* [1997] suggested that it provided effective control on overfitting while retaining the ability to discover complex features in the artificial data.

2.5 Other Methods

- **Simple, Naive, or Idiot's Bayes Network** – "Simple" is an approximation obtained by assuming independence between all variables and conditional independence given the disease. There is then a simple equation for the conditional probabilities. Although they work surprisingly well in many instances, this is not guaranteed. An equivalent one-layer (i.e., linear) ANN can be formulated.
- **Bayesian Reasoning and Decision Theory** - Bayesian networks model a joint probability distribution by factoring using the chain rule in probability theory. Although the models are very powerful when built, there exists presently no general machine learning methods for their construction. A considerable effort in this direction is needed. We have to divide the problem into independent partitions, called probabilistic similarity networks. It is inevitable that errors will occur when such large numbers of assessments are involved. An additional

drawback is that general probabilistic inference on Bayesian networks is NP-hard (NP-hard problem is at least as hard as or harder than any problem in NP, where NP means nondeterministic polynomial time), even for restricted networks.

- **Decision Trees** - Decision trees are domain-independent machine learning. They also have a problem in dealing with incomplete data. A piece of information can only be used if the appropriate question comes up when traversing the tree. Irrelevant questions cannot be avoided since the tree is always traversed in the same way starting from the root.

2.6 Evaluation of the Methodology

There are many ways to evaluate the performance of different classifiers, for example, the contingency table, the correct classification rate (CCR), the average squared error (ASE), the constant predictor (CP), the receiver operating characteristic (ROC) curve, transinformation content, etc.. The evaluation methods that have been used in this research are the contingency table, the correct classification rate, the average squared error and the constant predictor.

- **The Contingency Table** – For cross-classified data, the contingency tables (cross-tabulations) record the frequency for the values that fall into each possible combination of levels from two different factors. One of the factors is associated with the columns of the contingency table; the other factor is associated with the rows of the contingency table. From the contingency table, the operating characteristics (sensitivity, specificity, *etc.*) of the diagnostic tests can be easily calculated (refer to Appendix 2).

- **Correct Classification Rate (CCR) and Average Squared Error** – Correct classification rate is a measure of the ability of a classifier to correctly classify the test data set. Average squared error is an indicator of the error between the desired responses and the actual output of the classifier over the entire input space of interest. In the ANN training process, it is desired that the correct classification rate increases and the average squared error decreases as the number of epochs goes up.
- **The Constant Predictor (CP)** – Constant Predictor is a simple statistical benchmark that classifies all patterns as belonging to the class with the highest training set *a priori* probability. It is the minimum acceptable ANN performance.
- **The receiver operating characteristic (ROC)** – One measure of accuracy is the area under the curve of the receiver operating characteristic (ROC). Generally, ROC is a nonparametric measure of discrimination. The ROC measures the relative goodness of the set of predictions as a whole by comparing the predicted probability of each patient with that of all other patients. The computational approach to the ROC that employs the trapezoidal approximation to the area under the receiver operating characteristic curve for binary outcomes was first reported by Bamber [Bamber, 1975], and later in the medical literature by Hanley [Hanley and McNeil, 1982]. This was extended by Harrell [Harrell *et al.*, 1988] to continuous outcomes. Burke [1996] measured the prediction accuracy of ANNs and other statistical models by ROC curves for breast cancer survival. Although ROC curves make it fairly easy to compare the predictive power of different models, Buchman *et al.* [1994] pointed out that the value is limited. In particular, Buchman *et al.* noted that while ROC analysis may appear

to make models that predict a frequent outcome, such as survival in a surgical ICU, look good (as evidenced by a "large" area under the ROC curve), these models may in fact fail completely to identify a rare, but very important outcome such as death [Trigg, 1997]

- **Transinformation Content** – Transinformation refers to the amount of information transmitted through a predictive model and is a measure of the reduction in uncertainty, which can be attributed to using a particular model to classify each measurement vector. The concept was first developed by Shannon and Weaver [1949] and has since been simplified by Buttner [1982]. When Buchman *et al.* [1994] studied the performance difference between a multiple logistic regression equation and three separate ANNs in predicting the "chronicity" in a surgical ICU, he used this measure of information theory. A large value of transinformation content indicates that the input and output are closely interrelated. Zero means that the input and output are totally independent, that is, no useful information can be extracted from the model [Tong, Frize and Ennett, 1998].

2.7 Comparison between Neural Networks and Statistical Techniques

Neural networks have the potential to automatically discover complex relationships. There has been much interest in using neural networks in biomedical applications. Buchman *et al.* [1994] compared the logistic regression model and four connectionist models (including probabilistic neural networks) for the prediction of chronicity in a surgical intensive care unit. They found neural networks predicted more reliably than the statistical model regardless of the former's architecture.

However, there are not yet sufficient comparisons or theory to come to a firm conclusion about the utility of neural networks in biomedical data analysis. To date, comparison studies, e.g., those by Michie, Spiegelhalter, and Taylor [1994], have had mixed results, and Jefferson *et al.*'s [1995] complaint that many "successful" applications of neural networks are not compared against standard techniques appears to be justified. Recently, Burke [1996] and Plate *et al.* [1997] carried on this research respectively in breast cancer and tobacco & alcohol and cancer. They gave a positive evaluation of neural networks.

In Burke's [1996] research, the performance of ten models in predicting five- year breast cancer survival from the Patient Care Evaluation (PCE) data set were compared. These models included: pTNM Stages, Principal Components Analysis (PCA), pruned and shrunk Classification And Regression Tree (CART) [Breiman *et al.*, 1984], stepwise logistic regression, multilayer perceptron neural networks (backpropagation ANN, cascade correlation ANN, and conjugate gradient descent ANN) and non-multilayer perceptrons (probabilistic ANN [Specht, 1990], and, fuzzy ARTMAP ANN [Carpenter *et al.*, 1991]). Using the area under the curve of the receiver operating characteristic (ROC) as the measure of comparative accuracy, the accuracy of these models was (Table 2-1):

Table 2-1: Patient Care Evaluation 1983 Breast Cancer Data (5-year Survival Prediction, 54 Variables)

Prediction Model	Accuracy – the area under the curve of the receiver operating characteristic (ROC)
PTNM Stages	0.720
Principal Components Analysis	0.714
CART, pruned	0.753
CART, shrunk	0.762
Stepwise logistic Regression	0.776
Fuzzy ARTMAP ANN	0.738
Cascade correlation ANN	0.761
Conjugate gradient descent ANN	0.774
Probabilistic ANN	0.777
Backpropagation ANN	0.784

They suggested that logistic regression and the backpropagation ANN were the most accurate prediction models for predicting five-year breast cancer-specific survival. The ANNs were as good as the best traditional statistical models.

Plate *et al.* [1997] reported a study of various neural networks and statistical modeling techniques applied to an epidemiological data analysis problem - the relationship between tobacco, alcohol and cancer. In their experiment, they used null model, logistic regression, stepwise logistic regression, generalized additive models, classification trees, ordinary neural networks, bagged neural network with early stopping, and neural network with Bayesian regularization. Models were compared by test-set deviance in evaluation so as to avoid overfitting bias and keep the same criterion as was used for fitting. The result showed that neural networks with Bayesian estimation of regularization parameters (to control overfitting) performed consistently but only slightly better than logistic models when the data lacked complex relationships. But they performed markedly better than logistic models when complex

relationships were present. They suggested that neural networks have the potential to discover unanticipated complex features in the data and also demonstrated that MacKay's [1995] Bayesian neural network methodology provides effective control of overfitting while retaining the ability to discover complex features in the data.

In Trigg's master thesis [1997], she introduced weight-elimination into backpropagation neural networks to predict the "ventilation duration of 8 hours". She compared the results with those of Constant Predictor (CP) and Minimum Distance Classifier (MDC). CP is a statistical tool that classifies all cases as belonging to the output class with the highest probability. MDC is a classifier where each input pattern in the training and test sets is assigned to the output class with the highest estimated *a posteriori* probability. The performance improvement is shown below (Table 2-2).

Table 2-2: Comparison of Maximum Test Set Classification Rates Obtained Using the Best-performing Single and Double-layered ANNs to the Classification Performance Calculated for a CP and a MDC (Trigg, 1997)

Experiment	Max Test Set CCR	CP	Performance Improvement over CP	MDC	Performance Improvement over MDC
Best single-layered networks	90.5 ± 1.2%	71.1%	19.4%	86.1%	4.4%
Best double-layered networks	91.8 ± 1.15%	71.1%	20.7%	86.1%	5.7%

3. OVERVIEW OF NICU DATABASES AND RESEARCH OBJECTIVES REGARDING THEIR ANALYSIS

3.1 Type of Data in NICU Databases

The Canadian NICU Network has provided us two different versions of NICU databases in March 1999 (version 1) and August 1999 (version 2) respectively. The main difference between version 1 and version 2 is that version 1 contains the data collected by 5 Canadian NICU hospitals while version 2 contains the data from 17 Canadian NICU hospitals during the period – January 8, 1996 to October 31, 1997. The medical information in the NICU databases can be summed up in the following table (Table 3-1).

Table 3-1: The Medical Information in the NICU Databases

Admission Data Elements	Including demographic information such as birth-weight, Apgar scores, gestational age, case number and local centre code
Obstetric Characteristics	Including prenatal care, pregnancy information, intrapartum complications and treatments such as corticosteroids
Illness Severity Information (including data elements recorded on admission, day 3, 14 and 28)	SNAP score (items) CRIB score (items) NTISS score (62 items)
Discharge Data Elements	Including primary outcome variables: in-hospital death, major morbidity such as IntraVentricular Hemorrhage (IVH), BronchoPulmonary Dysplasia (BPD), necrotizing enterocolitis (NEC), nosocomial bacteremia, and, Retinopathy Of Prematurity (ROP)
Diagnoses and Conditions	Including all diagnoses recorded on the patient chart according to the International Classification of Diseases 9 (ICD9)
Disposition	Including locations to which patient was discharged and autopsy results in case of death
Resource Use	Including Length of Stay (LOS), transfusions, days of mechanical ventilation, parenteral nutrition and central catheter use

From the structural point of view, the databases (both version 1 and version 2) are composed of three sub-databases (Table 3-2 and Table 3-3): SNAP, FLAT, and NTISS. SNAP and NTISS include all the information about the patient measures while FLAT contains all the information about the patient outcomes.

Table 3-2: Source Information of the NICU databases (Version 1)

Database	File	# of cases	# of variables	Comment
SNAP	Snap_1	14997	78	All the Day1 (admission day), Day3, Day14 and Day28 data
FLAT	Flat_1	6917	129	Information about the outcomes
NTISS	Ntiss_1	14974	64	Categorical variables: Includes Day1, Day3, Day14 and Day28

Table 3-3: Source Information of the NICU databases (Version 2)

Database	File	# of cases	# of variables	Comment
SNAP	Snap1a.zip	10,000	83	All the Day1 (admission day) data
	Snap1b.zip	10,125	83	20,125 cases in total
	Snap3.zip	14,267	83	Day3 data
	Snap14.zip	9353	83	Day14 and Day28 data
FLAT	Flat1.zip	10,000	130	Information about the outcomes
	Flat2.zip	10,488	130	20,488 cases in total
NTISS	Ntiss_1.zip	43,603	66	Categorical variables: Includes Day1, Day3, Day14 and Day28

Since there is only a little difference in the structure between version 1 and version 2, let's take version 2 as an example to have a close look at the variables in each sub-database.

3.2.1 Variables in the Original SNAP Database (20125 cases, 83 variables)

The 83 variables in the original SNAP database (Version 2) are shown in Table 3-4 (The 51 variables in bold are needed to calculate the SNAP score, which will be talked about later.).

Table 3-4: Variables in the Original SNAP Database (Version 2)

DAY	**HBLOODP**	**HHEARTR**	**HRESPR**	**HSODIUM**
HPOTASS	**HHCO3**	**HHEMA**	**HPCO2**	HAPFIO2
LBLOODP	**LHEARTR**	**LTEMP**	**LSODIUM**	**LPOTASS**
LHCO3	**LHEMA**	**LSERUM**	LAPFIO2	**LPO2FIO2**
LPO2MAWP	LPO2PH	**LPO2PO2**	LPO2PCO2	**MAWPFIO2**
MAWPMAWP	MAWPPH	**MAWPPO2**	MAWPPCO2	**FIO2FIO2**
FIO2MAWP	FIO2PH	**FIO2PO2**	FIO2PCO2	**CBC1WBC**
CBC2WBC	**CBC3WBC**	**CBC1POL**	**CBC2POL**	**CBC3POL**
CBC1BAN	**CBC2BAN**	**CBC3BAN**	**CBC1PLT**	**CBC2PLT**
CBC3PLT	**SEIZURE**	**APNEA**	**GUAIAC**	**URINE**
WTG	NURPAT	MEDICUS	GRASP	PRN
FINISH	FINDATE	SCOREOLD	SCERR	IIERR
IIPEERR	***BTHWT***	APGAR5	SGA	URINE_OP
INDBILI	**HCALCTOT**	**HCALCION**	**HGLUC**	**HIDBILI**
HDBILI	**HBUN**	**HCREAT**	**LCALCTOT**	**LCALCION**
LGLUC	GESTAGE	SPI	SPII	SPPEII
SPIPE	HOSPITAL	CASELINK		

3.2.2 Variables in the Original NTISS Database (20488 cases, 130 variables)

The 66 variables in the original NTISS database (Version 2) are shown in Table 3-5

Table 3-5: Variables in the Original NTISS Database (Version 2)

DAY	ID	RESP	CPAP	MECVENT
VENTREL	HFVENT	SURF	INTUB	ECMO
NITOX	ROUTINE	NEW24HR	PERIPH	ARTERIAL
CENTVEN	ABX	ENT	PAREN	AMINOPH
NARCB	NACI	INDOM	ACID	STERIOD
CONVUL	KBIND	ERYTHRO	INSULIN	KDRIP
IVIG	RXOTHER	FREQ	ECG	WARMER
NONI_O2	ARTPRESS	CVP	CATHETER	QUANT
BLOODDR	TRANS	THORAC	DIALYSIS	CH_TUBE
OPERAT	CENTESIS	TUBE	GAVFED	AMINO
FATEMUL	PHOTO	PRBC	PLATELET	WBC
PATEXCHG	VOLEXCHG	VOLEXP	PRESS	PACE
CPR	FINISH	FINDATE	SCORE	CASELINK
HOSPITAL				

3.2.2 Variables in the Original FLAT Database (20488 cases, 130 variables)

The 130 variables in the original FLAT database (Version 2) are shown in Table 3-6. The variables in bold give the outcomes that we are interested in. Among these, "Duration of

Ventilation" (VENTDAY), "Mortality" (TBLDISC_=7 means non-survivor), and "Length of Stay" (if "DOND" exists, LOS=DOND-DOA; otherwise, LOS=DISDATE-DOA) are some very important outcomes.

Table 3-6: Variables in the Original FLAT Database (Version 2)

BTHWT	DOB	TOB	**DOA**	TOA
COMMENT	EPIGR	EPIEQ	SEX	RACE
BIRTHS	APGAR1	APGAR5	APGAR10	OBSEST
PEDEST	ADHEADC	HIFIO2	LOFIO2	WBD
ADTEMP	ADSTAT	TRANS	PAYOR	POSTAL_C
SHE	**IVHII**	**IVHIII**	**IPE**	**PVL**
STAGE	ZONE	PLUS	LEVENS	USDATE
VENTSIZE	ANT_CORT	DELIVERY	MAT_HYP	PRE_CARE
PRESENT	TBLPTRAN	**DOND**	TRANOTH	DAYO2
TBLWEEK3	V54	V55	V56	TBLDAY28
V58	V59	V60	WGT	V62
PDA	ISCHEM	RDS	PNEUMO	PNDATE
SEIZURE	CONGEN	**BPD**	LASTVENT	LASTO2
LAPSR	THORAC	RESER	ECMO	CRYO
OPMAJ	OPMIN	NECRO	NECRODAT	HIBILI
DATEBILI	**TBLDISC_**	AUTOPSY	DISWGT	V87
DISDATE	TOTHER	LIST	V91	THEOPHY
DIURETIC	MONITOR	GAVAGE	INCUB	CPAP
TRACH	GASTRO	IMV	R_OTHER	OSTOMY
COD	**VENTDAY**	CPAPDAY	O2DAY	NARCDAY
MORP	FERT	MEP	CODINE	METH
OPI	N_OTHER	ANDAY	DIAZ	LORA
MIDA	ASA	ACE	PHEN	TOPAN
A_OTHER	RELDAY	PANC	SUCC	VECURO
M_OTHER	BLDCULT	CSFCULT	READM	READMOBS
RETEMP	ID2	ALLORDER	BASE	CASELINK

3.2 Research Objectives

As it has been mentioned before, the currently used scoring systems are static, which means that they use the patient admission information only. Whether using time-varying data (up-to-data), as the patient's status changes over time, would add relevance and usefulness for

medical practitioners hasn't been studied yet. Therefore our long-term research goal is to establish the impact of time-varying data on the prediction of the patient outcomes.

Before investigating the time-varying characteristics, we need to have a good understanding of the relationship between non-time-varying data (admission day data only) and the patient outcomes. In this thesis, the artificial neural network (ANN) approaches were developed to estimate the neonatal ICU outcomes using the admission day data.

In summary, the objectives are:

1. Extend the ANN adult model (which was used to predict "duration of ventilation" in an adult ICU, developed by Trigg [1997]) to neonatal data; verify whether the model still works with a new type of data.
2. Study in depth the preprocessing of the neonatal data.
3. Develop software in C++ to see if it comes up with a more efficient way to perform ANN training and testing.

3.3 Selection of Databases for Prediction

As "Mortality", "Length Of Stay" and "Duration of Ventilation" are of great interest to the NICU physicians, the "Mortality" and "Duration of Ventilation" were selected as the research objects. Much effort was put in the study of "Duration of Ventilation" in order to compare the results with those from the adult ICU by Trigg [1997].

Based on the databases that we have, there are two sets of variables that we can use to predict these outcomes: using the variables in SNAP scoring system or using the variables in

NTISS scoring system. There are two versions of SNAP scoring systems, namely, SNAP and SNAP II (a parsimonious illness severity score for neonates).

Table 3-7: SNAP Score (Score for Neonatal Acute Physiology)
Ref: Score for Neonatal Acute Physiology: A Physiologic Severity Index For Neonatal Intensive Care
D. Richardson *et al.* Pediatrics Vol. 91 No. 3, 617-623, 1993

Parameter		Variable
Blood pressure	High	hbloodp
	Low	lbloodp
Heart rate	High	hheartr
	Low	lheartr
Respiratory rate		hrespr
Temperature C		ltemp
PO2		Lpo2po2
PO2/FiO2		See note 1.
PCO2		hpco2
Oxygenation index		mawp/Pao2*FiO2
Hematocrit	High	hhema
	Low	lhema
White blood cell count		lowest of the 3 cbc*wbc
Immature total ratio		See note 2
Absolute neutrophil count		See note 3.
Platelet count		lowest of cbc3plt
Blood urea nitrogen		Hbun
Creatinine		Hcreat
Urine output		Urine
Indirect bilirubin > 2 kg mg/dl or ≤ 2 kg mg/dl/kg		Hidbili
Direct bilirubin		Hdbili
Sodium	High	Hsodium
	Low	Lsodium
Potassium	High	Hpotass
	Low	Lpotass
Calcium (ions)	High	Hcalcion
	Low	Lcalcion
Calcium (total)	High	Hcalctot
	Low	Lcalctot
Glucose	High	Hgluc
	Low	Lgluc
Serum bicarbonate	High	hHCO3
	Low	lHCO3
Serum pH		Lserum
Apnea		Apnea
Stool Guaiac		Guaiac

Note

1. PO2/FiO2 (lowest) – record the lowest PaO2/FiO2 ratio from lpo2fio2, lpo2po2,mawpfio2, mawppo2, fio2fio2 and fio2po2,
2. Immature/total neutrophil ratio (high) = bands/(bands + polys)
3. Absolute neutrophil count (low)= the product of the lowest WBC count (including zeros) times the percent of polys + band.
4. The following equations give the definitions of the variables that require calculations:
   lowest pO2/FiO2 ratio = MIN(lpo2po2/lpo2fio2, mawppo2/mawpfio2, fio2po2/fio2fio2)
   highest oxygenation index = MAX(lpo2mawp/lpo2po2*lpo2fio2,
   mawpmawp/mawppo2*mawpfio2,
   fio2mawp/fio2po2*fio2fio2)/100
   lowest WBC = MIN(cbc1wbc, cbc2wbc, cbc3wbc)
   highest immature/total neutrophil ratio = MAX(cbc1ban/(cbc1ban+cbc1pol),
   cbc2ban/(cbc2ban+cbc2pol), cbc3ban/(cbc3ban+cbc3pol))
   lowest absolute neutrophil count = MIN(cbc1wbc*(cbc1pol+cbc1ban),
   cbc2wbc*(cbc2pol+cbc2ban), cbc3wbc*(cbc3pol+cbc3ban))
   lowest platelet count = MIN(cbc1plt, cbc2plt, cbc3plt)

Table 3-8: SNAP II Score (Score for Neonatal Acute Physiology II)

Parameter	Variable
Blood pressure Low	Lbloodp
Temperature C	Ltemp
PO2/FiO2	See note 1 of Table 4-2.
Urine output	Urine
Serum pH	Lserum
Seizure	Seizure

The numbers of variables in SNAP and SNAP II are 37 and 6 respectively (Table 3-7 and Table 3-8). Although as an objective scoring system, SNAP is valid for all birth weights and all major organ systems, it has too many variables. It is more suitable for research use instead of general clinical use. SNAP II was developed by Canadian NICU Network. It has only 6 variables that were extracted from SNAP using a statistical method: logistic regression. It is said that the discrimination ability of SNAP II score could approach SNAP and is better than another physiology based neonatal illness severity score (CRIB) (Lee, 1998).

In the NTISS database, since the variables "DAY", "ID", "FINISH", "FINDATE", "SCORE", "CASELINK", and, "HOSPITAL" are less related to the outcomes, they were excluded from further study. Because "PACE" is closely related to "Ventilation", it is also excluded from the prediction of "Duration of Ventilation". All other 68 variables were used as inputs.

It is expected that the effectiveness of SNAP and SNAP II score to predict patient outcomes using neural networks can be validated. Based on the assumption and experiences that neural networks can perform better than statistical models and a better scoring system is expected to be extracted from SNAP score using Backpropagation with weight-elimination techniques. Therefore, the following neural network models (shown in Table 3-9) were established.

Table 3-9: Neural Network Models

Scoring System	Input	Output
NTISS	68 variables in NTISS Scoring System	Duration of Ventilation <= or > 8 hours
		Duration of Ventilation <= or > 12 hours
		Duration of Ventilation <= or > 24 hours
SNAP	37 variables in SNAP Scoring System	Duration of Ventilation <= or > 8 hours
		Duration of Ventilation <= or > 12 hours
		Duration of Ventilation <= or > 24 hours
		Mortality
SNAP II	6 variables in SNAP II Scoring System	Duration of Ventilation <= or > 8 hours
		Duration of Ventilation <= or > 12 hours
		Duration of Ventilation <= or > 24 hours
		Mortality

As many NTISS variables (e.g. 'mecvent' which means mechanical ventilation, 'intub' which means intubation) have been found to be closely related to the output ('duration of

ventilation') after some preliminary experiments, the NISS database was not actually used in later research. More detailed explanation was given in the technical report.

Therefore, we will focus on the research based on SNAP and SNAP II scoring systems hereafter.

4. METHODOLOGY

4.1 ANN Model 1: Extending Adult ANN Model to NICU Database

4.1.1 ANN Model in Adult ICU by Trigg [1997]

A generalized feedforward backpropagation ANN architecture was constructed by Trigg [1997] in MATLAB in such a manner to allow a user to implement various combinations of the weight-elimination cost function, the standard backpropagation cost function as well as the regular data presentation techniques. The network employed hyperbolic tangent transfer functions. The cost function minimized during training in the standard backpropagation networks was the Sum of Squared Errors (SSE) between the target values and the actual network outputs. ASE values for experimental simulations were calculated only after training was complete.

The database Trigg used to validate this model was an adult ICU database from Doctor Everett Chalmers Hospital (DECH). The predicted outcome was whether the duration of ventilation of the patient is less than or equal to 8 hours (Vent8).

The database consists of 1491 patient records and 51 input variables (gender, age, admission diagnoses, location from which the patient was admitted, Glasgow Coma Score (GCS), death, days of concentrated nursing care, days in ICU, and physiological variables such as heart rate, partial pressure of oxygen, etc.).

In this well-collected database, there were a few incomplete records, i.e., some fields of these records were missing. Because ANN cannot deal with missing data, Trigg replaced the missing values (denoted by a value of –1 on the patients' ICU database record sheet) by 'normal' values given by the physician. The reason is "by using a data standardization (scaling) procedure that involves subtraction of the normal value from all continuous and integer-valued parameter values, all missing values can be converted to 0 inputs which will not have any impact on the ANN results. This procedure is beneficial in that it allows the neural network to make use of the valid information obtained from partially complete records" [Trigg, 1997].

In addition to the 51-input database, Trigg also derived some other databases. One database had 65 inputs generated by separating all continuous and integer-valued parameters as "high" (higher than normal) or "low" nodes (lower than normal). The other kind of database was obtained after separating the database by 'post-operation' (POSTOP) and 'non-post-operation' (NON-POSTOP) because they were suite different patient groups. Both single-layered and double-layered ANN architectures (with and without weight-elimination technique) were used to predict the output (duration of ventilation).

4.1.2 Extend Adult Model to Neonatal ICU Database

In order to investigate whether the adult model developed by Trigg could be applied to the neonates, similar techniques (ANN with and without weight-elimination) were used in a NICU database.

This database was provided by Canadian NICU Network in March 1999 (named NICU database version 1 in our research). It consists of information on 6,905 NICU patients from 5 of 17 Canadian hospitals. Different from the adult ICU database used by Trigg, this NICU database has a substantial number of incomplete records. Of the 6,905 records, 5,584 have missing information in some fields. Due to lack of simple and accurate method to deal with such a large amount missing information in ANNs, these incomplete records were not included in this work.

The SNAP II scoring system (6 input variables) was used to predict the same output as the adult model. Therefore the ANN classifier had 6 input variables and 1 output variable. The six inputs were LBLOODP (low blood pressure), LTEMP (low temperature), LSERUM (low serum pH), SEIZURE (presence of seizures) and URINE (urine output) and PO2/FiO2 ratio. The one output was whether a patient will require less than or equal to 8 hours of ventilation (VENT8).

Of the 1,321 complete records, 441 were set aside for testing, leaving 880 records for training, according to the pattern "train, train, test, train, train, test, ...". Each network was trained three times from different small random starting weights with and without weight-elimination (to control overfitting). The one with the best performance on the testing data was selected as "the model".

One characteristic of this approach is that only the complete records (19.1% of the database) were used for classification. Considering that the incomplete records may contain very important information, new methods will be needed to make full use of the large amount of missing information for a better prediction.

In order to have a good understanding of the database, database examination and data preprocessing (section 4.2) were done before new methods are proposed (section 4.3).

4.2 Database Examination and Data Preprocessing

Basically, the NICU database Version 2 has the same structure as Version 1 (refer to Table 3-2 and Table 3-3). Because Version 2 contains much more data than Version 1, it will be a more valuable database for further research. Therefore, in the following sections, we will focus on the processing of Version 2. Same procedures have been used to deal with Version 1.

The observed data in the NICU are recorded for each patient. The variables are divided according to their observed frequency and regularity as well as to their data types. We distinguish three kinds of data: continuously assessed quantitative data (e.g., high blood pressure), discontinuously assessed quantitative data (e.g., age), and qualitative data (e.g., name, sex) [Miksch et al., 1996]. To do the basic statistical analysis on the raw databases, a statistical software package (SPSS) was chosen. By means of SPSS, it is easy to find out the natural distribution, identify the missing information, remove the outliers, and analyze the correlation coefficients between the variables, etc..

At the beginning of the statistical study of the SNAP database, we found that: among the 20125 cases, 13 cases did not have "CASELINK" which means that their outputs are not available. They were excluded for further research.

4.2.1 Missing Values and Their Treatment

Missing values are widely spread in all of the three raw databases. According to the physician (Dr. Walker), there are several reasons that may result in their generation: 1) the patient did not need to have the medical procedures or some procedures were not available; 2) the physiological measurements were taken but the records were not recorded by the nurses.

4.2.1.1 Missing Values

At the first glimpse of the database, many numerical variables contain '-9' or '88' as their observations. Then we found that different symbols to mark the missing values are used in the three databases (SNAP, FlAT and NTISS). SNAP database mainly uses '-9' to mark the missing values while FLAT and NTISS mainly use '88'. '-9' indicates that the variable was missing or left null. For example, some values in 'urine' and 'ventday' are coded as '-9'. '88' implies that the variable was scored as 'unknown' or 'not applicable'. Some values in 'seh' (subependymal hemorrhage, also known as grade I intra-ventricular hemorrhage), 'ivhii' (grade II intra-ventricular hemorrhage, 'ivhiii' (grade III intra-ventricular hemorrhage), 'ipe' (intra-parenchymal hemorrhage, also known as grade IV intra-ventricular hemorrhage) and 'pvl' (peri-ventricular leukomalacia, a cellular injury adjacent to the ventricular region) are coded as '88'. Some fields are left blank to represent that there is no data entry with those variables.

In general, there are not many missing values in the FLAT database. For example, only 272 out of 20488 cases in 'VENTDAY' (Duration of Ventilation) are missing and only 268 out

of 'TBLDISC_' (Mortality) are missing. Since only a small portion of cases do not have outputs and it is impossible to use those cases with missing outputs as the training or testing samples in neural networks, the cases with missing outputs were not included in this study.

The situation with the SNAP database is totally different. There are not only a large number of missing values (e.g., there are more than 15,000 cases that have missing values in HCALCTOT, HDBILI, LCALCION, etc., which is ¾ of the database size) but there are also some obviously invalid values (e.g., -99 in CBC3POL, -933000 in URINE OUTPUT).

4.2.1.2 Replacing Missing Values by 'Normal' Value

Because of the existence of a large number of missing values, it is hard to obtain enough complete records to train ANN. Only 1321 out of 6905 cases in version 1 are complete records if the SNAP II scoring system (6 input variables) is used. When the SNAP scoring system (37 input variables) is used, even fewer complete records can be found. Besides, the incomplete records may contain very important information. In the literature, there are several methods to convert the incomplete records to complete ones. For example, replacing the missing values by the 'mean', 'median' or 'normal' value. Since good results were obtained by Trigg [1997] with an adult ICU database by assigning the 'normal' values to the missing values, 'normal'-value method was adopted in our experiments.

The selection of 'normal' values was based on the following SNAP scoring form (Table 4-1) provided by Dr. Walker.

Table 4-1: SNAP Scoring Form

Parameter		Level 0	Level 1	Level 3	Level 5
Blood Pressure (mean)	High	<66	66-80	81-100	>=100
	Low	>35	30-35	20-29	<20
Heart Rate	High	<180	180-200	201-250	>250
	Low	>100	80-100	40-79	<40
Respiratory Rate	High	<60	60-100	>100	---
Temperature (°F)	Low	>96	95-96	92-94.9	<92
pO2	Low	>65	50-65	0-50	<30
pO2/FiO2 Ratio	Low	>3.5	2.5-3.5	0.3-2.49	<0.3
pCO2	High	<50	50-65	66-90	>90
Oxygenation Index	High	<0.07	0.07-0.2	0.21-0.40	>0.40
Hematocrit	High	<66	66-70	>70	---
	Low	>35	30-35	20-29	<20
White Blood Count	Low	>5.0	2.0-5.0	<2.0	---
Immature/Total Ratio	High	<0.21	>=0.21	---	---
Abosulute Neutrophil Count	Low	>999	500-999	<500	---
Platelet Count	Low	>100	30-100	0-29	---
Blood Urea Nitrogen	High	<40	40-80	>80	---
Creatinine (mg/dl)	High	<1.2	1.2-2.4	2.5-4.0	>4.0
Urine Output (cc/kg/h)	Low	>0.9	0.5-0.9	0.1-0.49	<0.1
Indirect Bilirubin (bili/kg for BW>2kg)[1]	High	<15	15-20	>20	---
(bili/kg for BW<2kg)[1]	High	<5	5-10	>10	---
Direct Bilirubin	High	<2.0	>2.0	---	---
Sodium	High	<150	150-160	160-180	>180
	Low	>130	120-130	<120	---
Potassium	High	<6.6	6.6-7.5	7.6-9.0	>9.0
	Low	>2.9	2.0-2.9	<2.0	---
Calcium [Total][2]	High	<12	>12	---	---
	Low	>6.9	5.0-6.9	<5.0	---
Calcium [Ionized][2]	High	<1.4	>1.4	---	---
	Low	>1.0	0.8-1.0	<0.8	---
Glucose (or Reagent strip)	High	<150	150-250	>250	---
	Low	>40	30-40	<30	---
Serum Bicarbonate	High	<33	>33	---	---
	Low	>15	11-15	<10	---
Serum pH	Low	>7.30	7.20-7.30	7.00-7.19	<7.10
Seizure		None	Single	Multiple	
Apnea		None	Responsive to Stimulation	Responsive to stimulation	Complete Apnea
Stool Guaiac		Negative	Positive	---	---

Note: 1,2 Mutually exclusive items

In the above SNAP scoring system table, 'level 0' means the normal range, 'level 1', 'level3' and 'level 5' represent different abnormal ranges in increasing order of severity.

Table 4-2 is a conversion form, where the parameters in the SNAP scoring system are mapped to the variables in the NICU SNAP database. Following the formula (based on the email of Dr. Herb Chan and Dr. Robin Walker) given in the table, 37 SNAP variables used for ANN models were calculated from 51 variables (variables in bold in Table 3-4 of Chapter 3) in the NICU SNAP database. The 'normal' values (chosen by Dr. Walker) of these 37 SNAP variables are listed.

Table 4-2: SNAP Variable Conversion & Normal Value Assignment

Parameter		Variable Name in NICU Database	Normal Value	No.
Blood Pressure	High	Hbloodp	66	1
	Low	Lbloodp	35	2
Heart Rate	High	Hheartr	180	3
	Low	Lheartr	100	4
Respiratory rate		Hrespr	60	5
Temperature C		Ltemp	35.5	6
PO2		lpo2po2	65	7
PO2/FiO2		Min of Three (lpo2po2/lpo2fio2, Mawppo2/mawpfio2, fio2po2/fio2fio2)	3.5	8
PCO2		hpco2	50	9
Oxygenation Index		Max of Three lpo2mawp/lpo2po2*lpo2fio2, Mawpmawp/mawppo2*mawpfio2, fio2mawp/fio2po2*fio2fio2	0.07	10
Hematocrit	High	Hhema	66	11
	Low	Lhema	35	12
White Blood Cell Count		Min of Three (cbc1wbc, cbc2wbc, cbc3wbc)	5.0	13
Immature Total Ratio		Max of Three cbc1ban/(cbc1ban + cbc1pol), cbc2ban/(cbc2ban + cbc2pol), cbc3ban/(cbc3ban + cbc3pol)	0.21	14
Absolute Neutrophil Count		Min of Three cbc1wbc*(cbc1pol+cbc1ban)*10, cbc2wbc*(cbc2pol+cbc2ban)*10, cbc3wbc*(cbc3pol+cbc3ban)*10	999	15
Platelet Count		Min of Three cbc1plt, cbc2plt, cbc3plt	100	16
Blood Urea Nitrogen		hbun	14.3	17
Creatinine		hcreat	106	18
Urine Output		urine	0.9*24*weight in kg	19
Indirect Bilirubin > 2kg mg/dl or <= 2kg mg/dl/kg		hidbili	If bthwt>2kg: 256.5 If bthwt<=2kg: 85.5*weight in kg	20
Direct Bilirubin		hdbili	34.2	21
Sodium	High	hsodium	150	22
	Low	lsodium	130	23
Potassium	High	hpotass	6.6	24
	Low	lpotass	2.9	25
Calcium(ions)	High	hcalcion	1.4	26
	Low	lcalcion	1.0	27
Calcium(total)	High	hcalctot	3	28
	Low	lcalctot	1.725	29
Glucose	High	hgluc	8.325	30
	Low	lgluc	2.22	31
Serum Bicarbonate	High	hhco3	33	32
	Low	lhco3	15	33
Serum PH		lserum	7.30	34
Apnea		apnea	None (1)	35
Stool Guaiac		guaiac	Negative (1)	36
Seizure		seizure		37

4.2.2 Outliers and Their Treatment

In the raw NICU databases, there are many outliers, which might resulted from the human errors in the recording or typing process. Since the existence of outliers can heavily bias the performance of neural networks, we need to identify them and eliminate their influence. First, statistical analysis was done on the raw database and extremely low/high values were recognized. Then the outliers were picked according to the consultation with Dr. Walker.

4.2.2.1 Extremely Low/High Values

The frequency analysis of the raw SNAP database was done by SPSS. By investigating the results, we found that some values are extremely low or high. They might result from human errors. But whether they are outliers or not is still subject to further investigation. After studying the cases one by one, we list all the extreme values in Table 4-3.

Table 4-3: Extremely Low/High Values in SNAP Database

Variable	Extremely Low Value	Extremely High Value
LHEARTR	0 (19) << min 8 (1)	
HRESPR	0 (1051) << min 1 (46). Too many cases have min value 0 (1051) comparing to 1 (46).	300 (1), 420 (1), 6037 (1), 7637 (1) >> max 180
LPO2FIO2	0.21(1), 1.00(2), 5.00(113), 7.31(1), 7.36(1), 20.00(1) out of range 21-100	555 (5). Too many cases have max value 100 (715) comparing to 99 (1).
MAWPFIO2	5 (12) << min 21 (798)	Too many cases have max value 100 (560) comparing to 99 (1).
FIO2FIO2	5 (29) << min 21 (1054)	555 (1). Too many cases have max value 100 (832) comparing to 99 (1).
HPCO2	-9 (225) << min 4.44 (1)	
LPO2MAWP	0 (1220). Too many cases have min value 0 (1220) comparing to 0.55 (1).	
FIO2MAWP	0 (473) << min 2. Too many cases have min value 0 (473) comparing to 2 (1).	
HHEMA		2670 (1) >> max 534
LHEMA		210 (1) >> max 80
CBC1WBC		976 (1), 1013 (1), 1019 (1), 1152 (1), 1158 (1), 1225 (1), 1442 (1), 1493 (1), 1959 (1), 1973 (1) >> max 334
CBC2WBC		93.8 (1), 211 (1) >> max 73.7
CBC3WBC		76.9 (1) >> max 49.6
CBC1BAN	-8 (1).	182.93 (1), 208.33 (1), 1400 (1) >> max 80
CBC1POL	-953.00 (1).	177.33 (1), 200.00 (1), 220.78 (1), 300.00 (2), 400.00 (5), 606.99 (1), 2803.28 (1), 4400.00 (1) >> max 100
CBC2BAN		Not very extreme high: 69.23 (1) >> max 59
CBC3BAN		Not very extreme high: 70.20 >> max 58.60
CBC3POL	-99 (1).	
CBC1PLT		9999.00 (1), 10055 (1) >> max 4314
CBC2PLT		6449 (1) >> max 672
CBC3PLT	0.29 (1) << min 5.00	Not very: 563 (1) >> max 459
HBUN		136 (1), 203 (1), 367 (1) >> max 84
HCREAT		Not very: 525 91), 639 (1) >> max 480
URINE (cc):	-93300.00 (1), -9.0(10556), -8 (2).	Not very: 501.50 (1), 547 (1), 597.48 (1) >> max 437
HDBILI		Not very: 222 (1), 226 (1), 268 (1) >> max 176
LPOTASS	0.1 (1) << min 1.4	
HCALCION		Not very extreme high: 7.0 (1) >> max 5.7
LCALCION		9.0 (1) >> max 4.6
HCALCTOT		10.00 (1), 10.50 (1), 12.70 (1) >> max 4.2
LCALCTOT	Not very: 0.10 (1), 0.11 (1) << min 0.67	10.50 (1) >> max 4.2
LSERUM		43 (1), 721.00 (1) >> max 7.79 otherwise.

Note: A(B) means "The number of value A is B". For example, "HRESPR: 0 (1051) << min 1 (46)" means "1051 cases have value 0 in HRESPR, which is much less than 1. Most cases are more than 1 in HRESPR"; "HHEMA: 2670 (1) >> max 534" means "only 1 case have value 2670 in HHEMA, which is much more than 534. Most cases are less than 534 in HHEMA".

4.2.2.2 Outlier Examination

To determine the outliers, two techniques were used to identify them: 1) according to the suggestion of the NICU doctor; 2) using a statistical method and three standard deviations.

- **Identity Outliers According to the Suggestion of NICU Doctor**

According to Dr. Walker (August 24, 1999), some values are obviously invalid in practice while some can only be identified as suspicious values because there is no absolute range for some physical measurements (Table 4-4).

One limitation of this method is that the human factor has an inevitable influence on the data processing because in reality there is no clear cut-off value for most of the variables. For example, it is hard to say that high respiration rate above 180 is impossible although most patients do have high respiration rate of less than 180.

Table 4-4: Classification of Extreme Values in SNAP

Type	Variable	Value & Number
Errors: the minimum value of all the variables should be 0	HPCO2	-9 (225)
	CBC1BAN	-8 (1)
	CBC1POL	-953.00 (1)
	CBC3POL	-99 (1)
	URINE	-93300.00 (1), -9 (10556), -8 (2)
Errors: their ranges should be 21-100	LPO2FIO2	0.21 (1), 1.00 (2), 5.00 (113), 7.31 (1), 7.36 (1), 20.00 (1), 555 (5)
	MAWPFIO2	5 (12)
	FIO2FIO2	5 (29), 555 (1)
According to the doctor, the values in BOLD are outliers which should be excluded from further analysis. But since there is no absolute ranges for the variables, some of them are identified as suspicious values. Note: the number in the bracket represents the times that the value in front of it appears in the database.	HRESPR	**300 (1), 420 (1), 6037 (1), 7637 (1)** suspicious: 180
	HHEMA	**2670 (1)** suspicious: 534
	LHEMA	**210 (1)**
	CBC1WBC	**976 (1), 1013 (1), 1019 (1), 1152 (1), 1158 (1), 1225 (1), 1442 (1), 1493 (1), 1959 (1), 1973 (1)** suspicious: 334
	CBC2WBC	suspicious: 211
	CBC1BAN	**182.93 (1), 208.33 (1), 1400 (1)**
	CBC1POL	**2803.28 (1), 4400.00 (1)** suspicious: 177.33 (1), 200.00 (1), 220.78 (1), 300.00 (2), 400.00 (5), 606.99 (1)
	CBC1PLT	**9999.00 (1), 10055 (1)** suspicious: 4314
	CBC2PLT	**6449 (1)**
	CBC3PLT	**0.29 (1)**
	LPOTASS	**0.1 (1)**
	HCALCION *(range: <2.5)*	**2.52 (1), 2.57 (2), 2.70 (1), 2.83 (1), 2.90 (3), 3.00 (1), 3.10(1), 3.20 (1), 3.40 (1), 3.50 (1), 3.60 (1), 3.70 (1), 3.90 (1), 4.00 (1), 4.23 (1), 4.30 (1), 4.70 (1), 5.40 (1), 5.70 (1), 7.00 (1)**
	LCALCION *(range: <2.5)*	**2.57 (1), 2.70 (3), 2.80 (1), 2.83 (1), 2.90 (3), 3.00 (1), 3.10 (1), 3.20 (1), 3.50 (1), 4.30 (1), 4.60 (1), 9.00 (1)**
	HCALCTOT *(range: <3.5)*	**3.60 (1), 3.92 (1), 4.00 (1), 4.10 (1), 4.20 (1), 10.00 (1), 10.50 (1), 12.70 (1)**
	LCALCTOT *(range: <3.5)*	**3.60 (1), 4.00 (2), 4.20 (1), 10.50 (1).**
	LSERUM	**43 (1), 721.00 (1)** suspicious: 7.79

- **Identify Outliers Using Three Standard Deviations**

The alternative to avoid the human influence is to rely on the statistical methods. Based on the assumption that most of the outliers are out of range of three standard deviations, the

outliers can be identified by calculating the three standard deviations from the mean and standard deviation. The low/high boundaries of three standard deviations can be used as the cutting edges. Specifically, three methods based on three standard deviations were utilized.

A. Three standard deviations method (a)

In this approach, the three standard deviations were calculated from the mean and standard deviation with the original database (raw SNAP database) directly. The aim to use the original database directly is that the impact of human judgement on the outlier determination can be minimized.

The limitations of this method are: (1) The large number of obviously invalid values heavily affects the mean and standard deviation of some variables; some outliers are within the range of three standard deviations and could not be eliminated. Take the variable 'URINE' for example. '-93300' in urine output is an obviously invalid value (All the values in the databases should be positive in order to be valid). Because its absolute value is much larger than other values, the mean and standard deviation were highly biased. The result is that the outlier '-8' is still within the range of the three standard deviations. (2) Many valid values were cut off (e.g. high blood pressure 'HBLOODP', low blood pressure 'LBLOODP', etc.).

B. Three standard deviations method (b)

To solve the first problem in method (a), the absolutely invalid values (e.g. negative numbers) were replaced with 'missing' before calculating the three standard deviations. Then all

the values out of the three standard deviations were replaced with 'missing'. Thus the very large invalid numbers cannot bias the mean.

The limitation of this method is that many valid values were cut off (e.g. high blood pressure 'LBLOODP', low blood pressure 'LBLOODP', etc.).

C. Three standard deviations method (c)

One difference of this method from method (a) and (b) is that the outliers were not treated as 'missing'. Instead, we treat the values below the low boundary of the three standard deviations as the low boundary value and replace the values above the high boundary of the three standard deviations with the high boundary value.

The other improvement of this method is that it addressed the issue in method (a) and (b) where the valid values were cut off by the three standard deviations. By examining the distribution curve of all the variables, Dr. Walker suggested that some values should not be trimmed by three standard deviations. These special variables are: 'HBLOODP' (high blood pressure), 'LBLOODP' (low blood pressure), 'HHEARTR' (high heart rate), 'LHEARTR' (low heart rate), 'HDBILI' (high direct bilirubin), 'HIDBILI' (low direct bilirubin), 'HSODIUM' (high sodium), 'LSODIUM' (low sodium), 'HPOTASS' (high potassium), 'LPOTASS' (low potassium), 'HHCO3' (high serum bicarbonate), 'LHCO3' (high serum bicarbonate), 'APNEA' (apnea), 'GUAIAC' (guaiac) and 'SEIZURE' (presence of seizure). Therefore after replacing the absolutely invalid values with 'missing' and calculating the three standard deviations, the above variables were not replaced with the low/high boundary of three standard deviations.

The feature of this method is that the best valid substitutions for the values out of three standard deviations were found. Values below the low boundary were replaced with the low boundary. Values above the high boundary were replaced with the high boundary.

4.3 ANN Model 2: Including Outlier and Missing Value Processing

4.3.1 Database Generation for ANN

Three models dealing with the missing information and extreme low/high values are presented. The approaches to solve the problem of missing information are the same. 37 variables used in SNAP scoring system were calculated according to Table 4-3. The problem of incomplete records was solved by replacing the missing values with the 'normal' values suggested by Dr. Walker (Table 4-3). The methods to treat the extreme low/high values are somewhat different. They use three standard deviation method (a), (b) and (c) respectively. The procedures of model (a), (b) and (c) are shown in detail.

A. Model (a)

- Merge original database files snap1a.sav and snap1b.sav and sort cases by 'CASELINK'. Save as file snap1.sav (20125 cases, 83 variables).
- Delete cases without 'CASELINK' (13 cases). Save as file snap2.sav (20112 cases, 83 variables). This is the original database.
- Only keep 53 variables that are needed to calculate SNAP score variables. Calculate the range of three standard deviations from mean and standard deviation. Re-code all values

out of this range as 'SYSTEM MISSING'. Save as file snap3_1.sav (20112 cases, 53 variables).

- Calculate SNAP scoring variables (pO2/FiO2 ratio, oxygenation index, white blood cell count, immature total ratio, absolute neutrophil count, and platelet count). Save as file snap4_1.sav (20112 cases, 71 variables).
- Drop temporal variables, only keep 37 SNAP scoring variables, birth weight 'BTHWT' (which will be used to calculate 'normal' values of urine output 'URINE', high direct bilirubin 'HIDBILI' later) and 'CASELINK' (which will be used to merge the database SNAP and FLAT later). Save as file snapFinal_1.sav (20112 cases, 39 variables).

B. Model (b)

- Merge original database files snap1a.sav and snap1b.sav and sort cases by 'CASELINK'. Save as file snap1.sav (20125 cases, 83 variables).
- Delete cases without 'CASELINK' (13 cases). Save as file snap2.sav (20112 cases, 83 variables). This is the original database.
- Re-code obviously invalid values with 'MISSING', for example, negative values, blood gas with lowest PO2 'LPO2FIO2', blood gas with highest MAWP 'MAWPFIO2' and blood gas with highest FiO2 'FIO2FIO2' outside range (21-100). Save as file snap2_valid.sav (20112 cases, 83 variables).
- Only keep 53 variables that are needed to calculate SNAP score variables. Calculate the range of three standard deviations from mean and standard deviation. Re-code all

values out of this range as 'MISSING'. Save as file snap3_2.sav (20112 cases, 53 variables).

- Calculate SNAP scoring variables (pO2/FiO2 ratio, oxygenation index, white blood cell count, immature total ratio, absolute neutrophil count, and platelet count). Save as file snap4_2.sav (20112 cases, 71 variables).
- Drop temporal variables, only keep 37 SNAP scoring variables, birth weight 'BTHWT' (which will be used to calculate 'normal' values of urine output 'URINE', high direct bilirubin 'HIDBILI' later) and 'CASELINK' (which will be used to merge the database SNAP and FLAT later). Save as file snapFinal_2.sav (20112 cases, 39 variables).

C. Model (c)

- Merge original database files snap1a.sav and snap1b.sav and sort cases by 'CASELINK'. Save as file snap1.sav (20125 cases, 83 variables).
- Delete cases without 'CASELINK' (13 cases). Save as file snap2.sav (20112 cases, 83 variables). This is the original database.
- Re-code obviously invalid values with 'MISSING', for example, negative values, 'LPO2FIO2', 'MAWPFIO2' and 'FIO2FIO2' outside range (21-100). Save as file snap2_valid.sav (20112 cases, 83 variables).
- Only keep 53 variables that are needed to calculate SNAP score variables. Calculate the range of three standard deviations from mean and standard deviation. Only re-

code some values out of this range as low/high boundary of three standard deviations. The variables that should be trimmed include: 'HBLOODP', 'LBLOODP', 'HHEARTR', 'LHEARTR', 'HDBILI', 'HIDBILI', 'HSODIUM', 'LSODIUM', 'HPOTASS', 'LPOTASS', 'HHCO3', 'LHCO3', 'APNEA', 'GUAIAC' and 'SEIZURE'. Save as file snap3_3.sav (20112 cases, 53 variables).

- Calculate SNAP scoring variables (pO2/FiO2 ratio, oxygenation index, white blood cell count, immature total ratio, absolute neutrophil count, and platelet count). Save as file snap4_3.sav (20112 cases, 71 variables).
- Drop temporal variables, only keep 37 SNAP scoring variables, birth weight 'BTHWT' (which will be used to calculate 'normal' values of 'URINE', 'HIDBILI' later) and CASELINK (which will be used to merge the database SNAP and FLAT later). Save as file snapFinal_3.sav (20112 cases, 39 variables).

The general procedure to prepare the database for ANN modeling is given by Figure 4-1. The graphical description of model generation process in detail is shown in Figure 4-2 and Figure 4-3.

Merging the final database from SNAP processing (snapFinal_*.sav) with that from FLAT processing (flat4.sav) results in the model used for NN analysis.

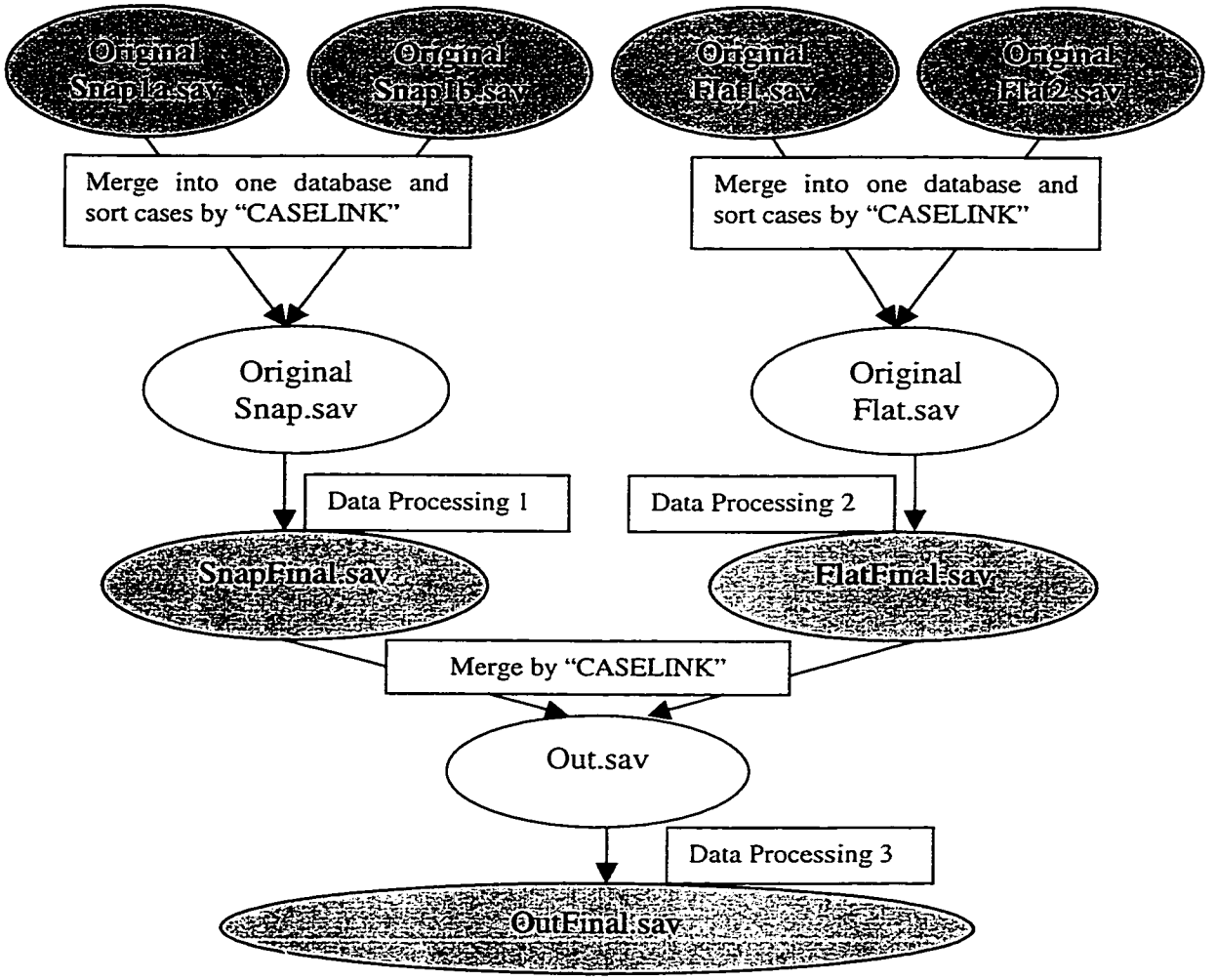

Figure 4-1: General Procedure of Database Preparation

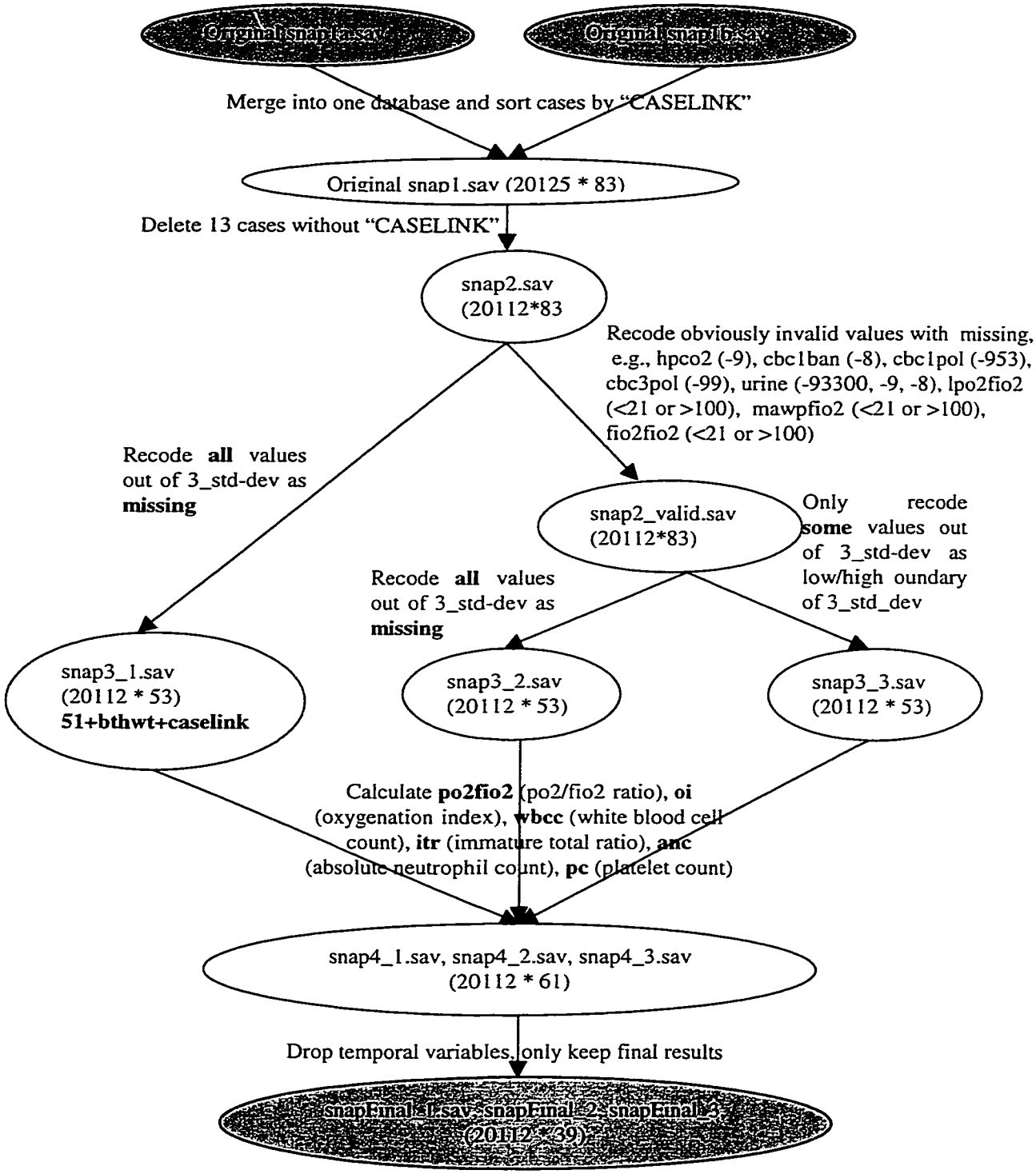

Figure 4-2: SNAP Database Processing (Detailed Data Processing 1 in Figure 4-1)

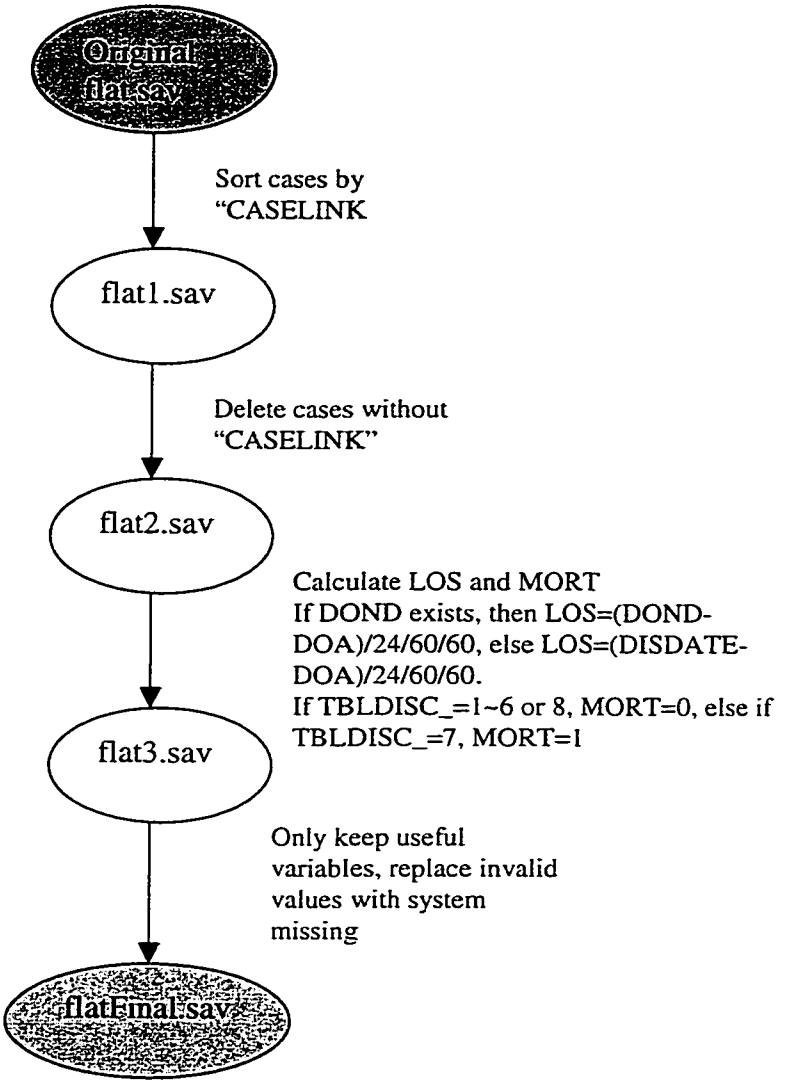

Figure 4-3: FLAT Database Processing (Detailed Data Processing 2 in Figure 4-1)

4.3.2 Experimental Simulation with ANN

After processing the missing information and the outliers, the database is trained by neural networks. Neural networks are in effect sigmoidal transformations of statistical regressions and hence are similar to a nonlinear regression approach to analyzing information in a database. The SNAP database contains the information on patient physiological measurement

and the FLAT database contains the information on patient outcomes. The inputs to the ANN model are patient measures (from the SNAP database) and the outputs are the patient outcomes (from the FLAT database).

Similar techniques used on the adult ICU database [Trigg, 1997] were applied to the NICU database and a comparison of the results were made. A feed-forward backpropagation ANN was adopted to investigate the relationship between the input parameters of SNAP scoring systems and the outcomes (i.e., duration of ventilation, mortality).

The training set to be employed in this experiment will consist of 2/3 of patient admissions in the database, while the remaining 1/3 will be used as a test set. Patients were assigned to the training and test sets randomly or sequentially in the pattern Train, Train, Test, Train, Train, Test, ...

To control the overfitting, the weight-elimination method [Weigend et al., 1991] used by Trigg [Trigg, 1997] was applied. The effectiveness of weight-elimination techniques in reducing over-fitting and improving the performance for both single-layer and double-layer ANNs were tested with the neonatal data. The performance of networks trained using weight elimination was compared to the performance of the same networks trained without weight elimination. Results were compared to the Constant Predictor and Correct Classification Rate was reported.

4.4 Developing ANN Model in C++

A generalized feedforward backpropagation ANN architecture was constructed in Borland C++.

There were several motives to develop this software. Firstly, in order to get the optimal performance of the test set, we need to continuously change the network parameters (e.g. number of hidden nodes, learning rate, momentum, etc.). If the files are written in MATLAB, whenever a parameter needs to be changed, we have to go back to modify the MATLAB files. It is very inconvenient. Secondly, training ANNs in MATLAB is very time-consuming. Its running speed is much slower than that of C++. Finally, the programs written in MATLAB cannot be used in commercial applications because everybody can gain access to the source codes.

Therefore, a feedforward ANN trained by backpropagation algorithm was realized in C++. The names of training and testing files, the number of samples in the training/testing sets, number of variables, the number of hidden nodes, the learning rate, the momentum and the maximum epoch can be adjusted when running the executable file. The source files won't be affected.

This program can be further improved by adding more features. For example, we can add some codes to allow the changes of average squared error and correct classification rate with time shown in graphic mode. This could be an interesting subject in future research.

Figure 4-4 shows the flow chart of this program. The source codes are included in a separate technical report and not attached here due to proprietary reasons. They are available only with the permission of MIRG's directors.

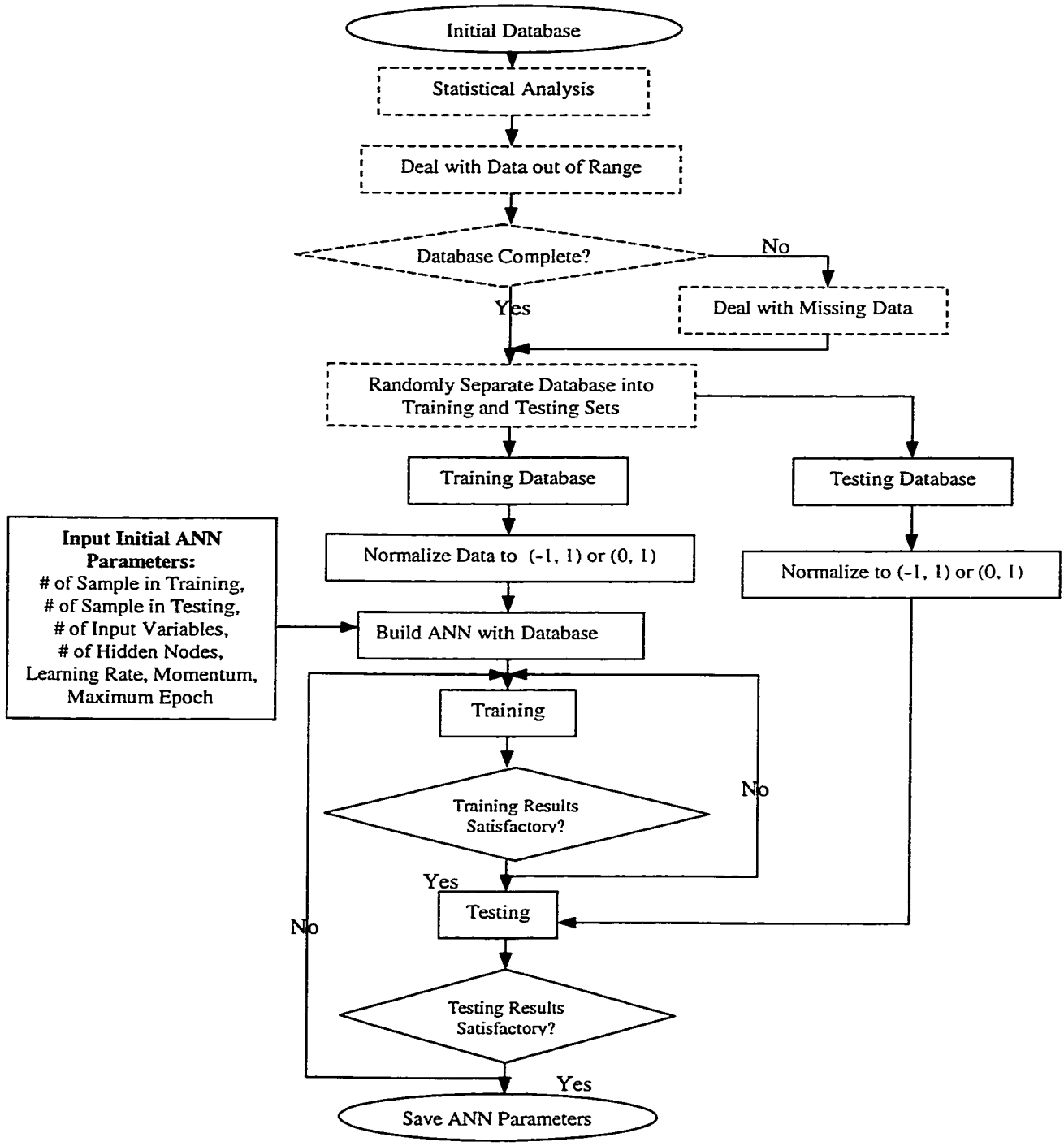

Figure 4-4: Flow Chart (Procedures in Dashed Blocks were Realized in SPSS, Procedures in Solid Blocks were Realized in C++)

5. RESULTS AND DISCUSSION

5.1 Comparison of Adult Model and Neonatal Model 1

Trigg explored the predictive capacity of ANNs to estimate whether a patient will require less than or equal to 8 hours of ventilation (VENT8). The best results obtained from the 'POSTOP' databases were reported in table 5-1 [Trigg, 1997]. Maximum correct classification rate (CCR), the CCR for the constant predictor (CP), performance improvement over CP, the CCR for the minimum distance classifier (MDC) and performance improvement over MDC, the maximum CCR with and without weight-elimination (WE) for the test set are reported.

Table 5-1: Comparison of maximum test set classification rates obtained using the best-performing single and double-layered ANNs (slregwte and dlregwte) to the classification performance calculated for a CP and a MDC. (Also listed in the chart are the maximum test set classification rates achieved by Buskard [1994] in preliminary experiments to estimate the duration of artificial ventilation for DECH ICU patients.)

Experiment	Max Test Set CCR	CP	Performance Improvement over CP	MDC	Performance Improvement over MDC	w/o WE	with WE
best single-layered networks							
slregwte(51:1)	90.5±1.2%	71.1%	19.4%	86.1%	4.4%	88.8%	90.5%
slhlwte(65:1)	89.5±0.85%	71.1%	18.4%	87.1%	2.4%	88.1%	89.5%
best double-layered networks							
dlregwte(51:1)	91.8±1.15%	71.1%	20.7%	86.1%	5.7%	90.5%	91.8%
dlhlwte(65:1)	91.5±1.0%	71.1%	20.4%	87.1%	4.4%	88.4%	91.5%
Buskard et al.'s 12-category estimation (41:5:12)	69.0%	66.5%	2.5%	66.3%	2.7%	N/A	N/A

Table 5-1 shows that both single-layered and double-layered ANN gave better classification results than CP, MDC and Buskard et al.'s 12-category estimation in predicting VENT8 in DECH ICU database. From Trigg's results, the double-layered ANNs performed slightly better than the single-layered ANNs, the ANNs with weight-elimination performed better than those without weight-elimination. Trigg successfully verified that ANN weight-elimination is an efficient way to eliminate over-fitting and improve network performance.

What needs to be mentioned here is that Trigg's work was further improved by Ennett [Ennett, 1998] and Ho [Ho, 1998]. Using a one-layered ANN with weight elimination technique, six variables (heart rate, respiratory rate, fraction of inspired Oxygen, partial pressure of Oxygen in the blood, arterial pH, and glasgow coma score) with the largest weights were extracted from 51 variables [Trigg, 1997]. An ANN with these six variables was used to predict the output and this simplified network generally presented a better performance in terms of CCR, ASE and training time [Ennett, 1998; Ho, 1998]. This result will be shown later.

Using six input variables chosen by the SNAP II scoring system and deleting all the records with missing information, similar results were obtained from the NICU database (Version 1) with ANN.

Table 5-2: Contingency Table: Predicting Duration Of Ventilation – NICU Model 1 (<=8hrs or >8hrs, when lr=5e-4, momentum=0.9)

	Without Weight-elimination		**With Weight-elimination**	
Training Set (apriori probabilities=63.45 % and 36.54%)	True Positive #	False Positive #	True Positive #	False Positive #
	559	**8**	**559**	**20**
	False Negative #	True Negative #	False Negative #	True Negative #
	0	**314**	**0**	**302**
Testing Set (apriori probabilities=66.36 % and 33.64%)	True Positive #	False Positive #	True Positive #	False Positive #
	292	**28**	**292**	**23**
	False Negative #	True Negative #	False Negative #	True Negative #
	0	**120**	**0**	**125**

Higher correct classification rate was achieved by using weight-elimination technique (Table 5-2). But no difference in performance has been seen between the single-layered ANN and the double-layered ANN.

Training and testing curves of the double-layered ANN trained by backpropagation algorithm (DLBP) with and without weight-elimination technique are shown in Figure 5-1 and 5-2 respectively. The decrease of the average squared error (ASE) and the increase of correct classification rate (CCR) with time (epoch), and, the slight improvement with weight-elimination over without weight-elimination can be seen clearly.

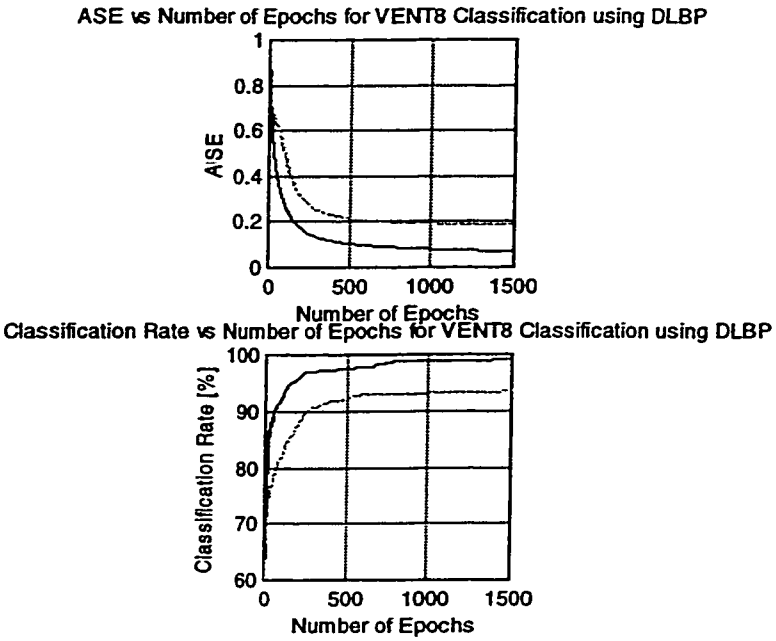

Figure 5-1: ASE and CCR for Ventilation<=8hr or >8hr with Weight-elimination (Solid line for Training, Dashed line for Testing)

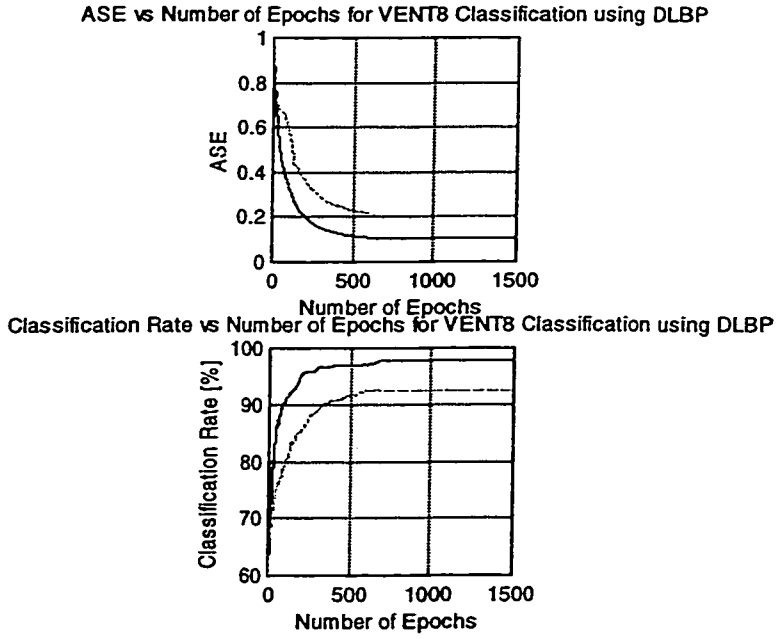

Figure 5-2: ASE and CCR for Ventilation<=8hr or >8hr w/o Weight-elimination (Solid line for Training, Dashed line for Testing)

Putting the results of double-layered ANN from DECH adult ICU (AICU) and the NICU database together, we can see good classification performance were achieved in all the three cases. Table 5-3 lists the number of variables, maximum correct classification rate (CCR) for the test set, average squared error (ASE), number of epochs and the CCR of the constant predictor (CP) in these models.

Table 5-3: Predicting Duration of Ventilation (<=8 hrs or >8hrs) in NICU & AICU by ANNs

	# of Variables	MAX Test Set CCR	ASE of MAX Test Set CCR	# of Epochs	CP
AICU with WE (by Trigg)	51	91.8%	0.35	2071	71%
AICU w/o WE (by Trigg)		90.5%	0.35	942	
AICU w/o WE (by Ennett and Ho)		90.5%	0.32	130	
NICU with WE	6	92.7%	0.13	540	61%
NICU w/o WE		91%	0.2	650	

Seen from above results, the ANNs estimate consistently better than the Constant Predictor (CP, a simple statistical benchmark that classifies all patterns as belonging to the class with the highest a priori probability in the training set) in terms of CCR and ASE. The performance of the model without weight-elimination is further improved by controlling overfitting with weight-elimination (WE).

The results show that the adult model could be successfully applied to neonates with the current database used. The results also show that ANNs can be useful tools in predicting ventilation based on the admission day SNAP II variables.

In the adult ICU there is an interesting result: after extracting the variables with large weights with the weight-elimination technique, the performance is significantly improved by

using only a small number of input variables (6 variables instead of the 51 original variables). Similar investigation was carried out in the neonatal ICU. We intended to discover whether the same variables as those in the SNAP II scoring system could be extracted from the 37 SNAP variables by means of the weight-elimination technique. But, even with the NICU databases Version 2 (which has about 20,000 cases), enough cases couldn't be obtained to train the ANNs after deleting all the incomplete records.

5.2 Results from Model 2 - (a) (b) (c)

For Model 2-(a), 2-(b) and 2-(c), two scoring systems, SNAP (37 input variables) and SNAP II (6 input variables), have been used to predict "Mortality" and "Duration of Ventilation" (<=8 hours, <=12 hours, <=24 hours, respectively). The common characteristic of these models is that the missing fields were replaced with the 'normal' values suggested by the physician because Trigg [1997] successfully validated this approach in an adult ICU database.

However, in the NICU database it seems that this approach did not work as well as we had expected. These models performed as the constant predictors no matter what the model architecture, learning rate, momentum and the weight-elimination parameter were chosen. They predicted all the cases as belonging to the class with higher apriori probability. Take the prediction of 'duration of ventilation <= 8 hours' (Vent8) for example (Table 5-4), the neural networks always classified all the patients (both duration of ventilation <= 8 hours and > 8 hours) as 'duration of ventilation <= 8 hours'. The reason is that most patients had 'duration of

ventilation' less than or equal to 8 hours (89.8% for the whole database, 89.4 for the training set and 91.5% for the testing set).

Table 5-4: Contingency Table: Predicting Duration of Ventilation (<=8hrs or >8hrs) – NICU Model 2

	Without Weight-elimination		With Weight-elimination	
Training Set (apriori probabilities=10.6% and 89.4%)	True Positive #	False Positive #	True Positive #	False Positive #
	11842	**1405**	**11842**	**1405**
	False Negative #	True Negative #	False Negative #	True Negative #
	0	**0**	**0**	**0**
Testing Set (apriori probabilities=9.5% and 91.5%)	True Positive #	False Positive #	True Positive #	False Positive #
	5997	**627**	**5997**	**627**
	False Negative #	True Negative #	False Negative #	True Negative #
	0	**0**	**0**	**0**

Similar results were obtained when predicting the mortality (Table 5-5). The neural networks always classified all the patients (both the survivors and non-survivors) as survivors though the real mortality rate is 3.78% (752 out of 19871 patients died).

Table 5-5: Contingency Table: Predicting Mortality – NICU Model 2

	Without Weight-elimination		With Weight-elimination	
Training Set (apriori probabilities=3.77% and 96.23%)	True Positive #	False Positive #	True Positive #	False Positive #
	12746	**500**	**12746**	**500**
	False Negative #	True Negative #	False Negative #	True Negative #
	0	**0**	**0**	**0**
Testing Set (apriori probabilities=3.80% and 96.20%)	True Positive #	False Positive #	True Positive #	False Positive #
	6371	**252**	**6371**	**252**
	False Negative #	True Negative #	False Negative #	True Negative #
	0	**0**	**0**	**0**

Note: There are 19871 cases in total: 752 cases have "MORTALITY=1"; 19117 cases have "MORTALITY=0"; two cases have missing information in "MORTALITY".

Though many experiments have been carried out, no effective solution has been found yet. One possible reason of this problem is that the artificial manipulation on the large amount of missing information might stabilize the neural networks to an average point of the system instead

of the optimal point. The missing information problem is far more complicated than originally anticipated with the neonatal data. More investigation is needed to solve this challenging problem.

5.3 Statistical Comparison

Are the two patterns (Vent8=1 and Vent8=0 or mortality=1 and mortality=0) separable? The results from model 2- (a), (b) and (c) bring up such a question.

Before wrapping up the whole session of data processing and modeling, let's have a look at some results from the statistics. The frequency analysis was used to compare the data of the two patterns in general. Table 5-6 compares the six SNAP II variables in two class (Vent8=1 and Vent8=0) respectively in terms of mean, median, etc.. The numbers in bold emphasize a large difference of the variable values in different classes.

Table 5-6: Statistical Comparison of Two Classes (Vent8=1 & Vent8=0)

		# of Valid	# of Missing	Percentage of Missing	Mean	Median	Minimum	Maximum
Vent8=1 (Duration of Ventilation > 8 hours) 2032 cases	LBLOODP	2007	25	1.23%	**30.05**	**28**	10	80
	LTEMP	2030	2	0.10%	36.02	36.2	21	44
	PO2/FIO2	1683	349	17.17%	**1.53**	**1.32**	0	8.05
	URINE	1684	348	17.12%	**28.3**	**17.55**	0	547
	LSERUM	1949	83	4.08%	7.3	7.32	4.4	7.7
	SEIZURE	2031	1	0.05%	1.03	1	1	3
Vent8=0 (Duration of Ventilation <= 8 hours) 17839 cases	LBLOODP	14673	3166	17.75%	**39.1**	**38**	0	105
	LTEMP	17764	75	0.04%	36.42	36.5	20	45
	PO2/FIO2	5439	12400	69.51%	**2.01**	**1.76**	0.02	12.42
	URINE	7737	10102	56.63%	**48.8**	**36**	0	597.48
	LSERUM	10001	7838	43.94%	7.39	7.33	6.4	721
	SEIZURE	17837	2	0.01%	1.04	1	1	3

It can be seen that compared to the patients with a duration of ventilation less than or equal to 8 hours, the patients with a duration of ventilation more than 8 hours tend to have lower LBLOODP (low blood pressure), lower LTEMP (low temperature), lower PO2/FIO2 (ratio of partial Oxygen and fraction of inspired Oxygen), lower URINE (urine output), lower LSERUM (low serum pH), and higher SEIZURE (seizure) in average. The difference in mean and median values of LBLOODP, PO2/FIO2 and URINE in the two classes (Vent8=1 and Vent8=0) seems to be a factor that should be studied in-depth in future work.

Similar results are seen when 'mortality' was considered (Table 5-7). But why the difference is not enough for neural network model 2 - (a), (b) and (c) to separate long ventilation hours and short ventilation hours, survivor and non-survivor, still needs to be further investigated. A good explanation to this question might be very helpful to explore the intrinsic relationship between the input data and the outcomes.

Table 5-7: Statistical Comparison of Two Classes (Mortality=1 & Mortality=0)

		# of Valid	# of Missing	Percentage of Missing	Mean	Median	Minimum	Maximum
Mortality=1 (non-survival) 752 cases	LBLOODP	709	43	5.72%	**30.35**	**28**	0	84
	LTEMP	722	30	3.99%	35.80	36	29.5	38.4
	PO2/FIO2	600	152	20.22%	**1.12**	**0.76**	0.03	5.76
	URINE	610	142	18.88%	**21.10**	**10**	0	227
	LSERUM	684	68	9.04%	7.21	7.26	6.4	7.58
	SEIZURE	751	1	0.13%	1.17	1	1	3
Mortality=0 (survival) 19117 cases	LBLOODP	15969	3148	16.47%	**38.35**	**37**	0	105
	LTEMP	19070	47	0.25%	36.41	36.5	34.4	38.3
	PO2/FIO2	6521	12596	65.89%	**1.95**	**1.71**	0	7.79
	URINE	8809	10308	53.92%	**45.63**	**33**	0	187.21
	LSERUM	11264	7853	41.08%	7.33	7.33	4.4	26.88
	SEIZURE	19115	2	0.01%	1.03	1	1	3

Note: There are 19871 cases in total: 752 cases have "MORTALITY=1"; 19117 cases have "MORTALITY=0"; two cases have missing information in "MORTALITY" (MORTALITY=SYSTEM MISSING).

5.4 Validation of ANN Model in C++

In addition to the NICU research, some work has been done in developing an ANN model with C++. To verify the performance of this tool, we used Trigg's 'POSTOP' database from DECH ICU as a benchmark.

As the weight-elimination technique has not been added yet to this model, only the performance of standard ANN trained with BP were compared [Table 5-8] in terms of the maximum CCR for the test set, the number of epochs and the time used.

Table 5-8: Performance Comparison of ANN Models Implemented with MATLAB and C++ When Predicting Ventilation (<=8 hrs) in AICU

	# of input Variables	MAX Test Set CCR	# of Epochs	Time (minute)	CP
AICU w/o WE (MATLA B)	51	90.5%	942	20	71%
AICU w/o WE (C++)		88.2%	600	6	

From the table above, we can see that comparing with the experiments using MATLAB as implemented by Trigg, the new program obtains a comparable performance but runs more than three times faster. The advantage in speed of this model can be very useful when dealing with some large databases and a large series of experiments. But further improvement on this model needs to be done to provide more features and a better performance.

6. CONCLUSION AND FUTURE WORK

6.1 Conclusion

Predicting death, ventilation, length of stay, other complications (e.g., lung disease, brain damage) could be helpful for physicians. The first contribution of this thesis is that it successfully validated the use of neural networks in predicting ventilation (<=8hours or >8hours) with SNAP II scoring system in a NICU database of about 1000 patients without missing value cases. The effectiveness of weight-elimination in over-fitting control was also validated. By comparing the result from an AICU database, it was verified that the adult model could be applied to the neonates under certain circumstances (model 1).

The second contribution is that the experiments identified a critical and challenging problem in improving the performance of neural networks, which is missing information. The missing information problem is very common in medical databases. No effective approach has been proposed and validated to make effective use of this kind of information so far. Therefore it is a challenging problem. From the experiments, it can be seen that using three standard deviations to eliminate the outliers and replacing the missing values with NORMAL values were not able to improve the performance of the neural networks. To solve the missing information

problem, much effort and many other methods are needed. For example, here are some potential approaches that could be attempted:

- replace the missing values with the mean or median value of the variable;
- use Backpropagation with more than one hidden layer;
- use a different architecture (Radial Basis Function);
- estimate the missing value by training the ANNs to make up the missing information, e.g. Expectation-Maximization algorithm;
- design a new architecture of NN that can process incomplete records.

The third contribution is the implementation of a backpropagation neural network model with C++. The training/testing file, number of samples in the training/testing sets, number of variables, number of hidden nodes, learning rate, momentum and maximum epoch can be adjusted outside the source file. Comparing with the model in MATLAB, this model runs three to four times faster. It has been tested with one adult ICU database. One shortcoming of this program is that the change of average squared error and correct classification rate in graphic mode has not been added yet. The source code has been provided to MIRG in a technical report.

6.2 Future Work

With such a large NICU database, lots of interesting research could be carried on. Here, some potential projects are listed.

(1) Temporal Analysis

As mentioned earlier, our ultimate objective is to generalize the approaches for use in a multitude of medical settings, and to provide a more up-to-date status of the patient than using only static "admission data". To realize this goal, the static system needs to be moved to a dynamic system (which uses data acquired at different points in time during the patient's stay in the NICU) after successfully validating the usefulness of ANN in predicting "Mortality", "Duration of Ventilation" and "Length of Stay".

Temporal analysis can be analyzed with the NTISS and SNAP scoring systems to study the influence of time on the patient outcomes and resource utilization. NTISS and SNAP scoring systems keep a record of patients' test data on Day 1, Day 3, Day 14 and Day 28. By comparing the performance of separate models (built on Day 3, Day 14 and Day 28 respectively) and mixed model (including all the patient data), it can be investigated whether the additional data (collected on Day 3, Day 14 and Day 28) provides more information for the predictions.

(2) Estimation of Rare Disease

With a large database, the relationship between the input parameters and several other outcomes that occur more rarely can be studied using a similar approach. Some examples of these outcomes are: in-hospital death, IVH (intraventricular hemorrhage), BPD (bronchopulmonary dysplasia), necrotizing enterocolitis, nosocomial bacteremia, retinopathy of prematurity, blood product transfusions, days of parenteral nutrition, days of central venous catheter use, receipt of antenatal corticosteroid, daily nurse/patient numbers, acuity of patients.

(3) Investigation of the Effect of Territory

The NICU data provided by Canadian NICU Network was collected from 17 Canadian hospitals. Will the site of the hospital (territory) or the services/procedures in the hospital affect the patient outcomes in general? One method is to separate the database by institution and compare the network performance of the sub-databases.

(4) Implementation and Testing of the Prototype

Once an index with fewer input variables than the previous scoring systems (SNAP, NTISS) is deduced with the weight-elimination technique, the prototype of a decision support tool with a user-friendly interface can then be developed to aid the physicians in making a diagnosis or to select their patient management strategy.

7. REFERENCE

[Ahmad and Tresp, 1993]
S. Ahmad and V. Tresp, **Some Solutions to the Missing Feature Problem in Vision.** In S. J. Hanson, J.D. Cowan and C.L. Giles, (Eds.), *Advances in Neural Information Processing Systems*, Vol. 5, San Mateo, CA: Morgan Kaufmann.

[Barnard, 1992]
E. Barnard, **Optimization for Training Neural Nets**, *IEEE Trans. Neural Net.*, Vol.3, No.2, p232-240: 1992.

[Battiti, 1992]
R. Battiti, **First- and Second Order Methods for Learning: Between Steepest Descent and Newton's Method**, *Neural Computation*, Vol.4, No.2, p141-166: 1992.

[Baxt, 1990]
W.G. Baxt, **Use of an Artificial Neural Network for Data Analysis in Clinical Decision-Making: The Diagnosis of Acute Coronary Occlusion**, *Neural Computation*, 2, p480-489: 1990.

[Baxt, 1991]
W.G. Baxt, **Use of an Artificial Neural Network for the Diagnosis of Myocardial Infarction**, *Annals of Internal Medicine*, 115, p843-844: 1991.

[Baxt, 1992a]
W.G. Baxt, **Analysis of the Clinical Variables Driving Decision in an Artificial Neural Network Trained to Identify the Presence of Myocardial Infarction**, *Annals of Emergency Medicine*, 21(12), p1439-1444: December, 1992.

[Baxt, 1992b]
W.G. Baxt, **Improving the Accuracy of an Artificial Neural Network Using Multiple Differently Trained Networks**, *Neural Computation*, 4, p772-780: 1992.

[Baxt, 1994]
W.G. Baxt, Complexity, **Chaos and Human Physiology: the Justification for Non-Linear Neural Computational Analysis**, *Cancer letters*, 77, p85-93: 1994.

[Bishop, 1995]
C. M. Bishop, **Neural Networks for Pattern Recognition**, Clarendon Press, Oxford: 1995.

[Brandt, 1997]

R. Brandt, **Application of Mixed Models to Problems in Medical Informatics**, J.S.M. Anaheim, Calif., 1997.

[Breiman *et al.*, 1984]
Breiman L, Friedman JH, Olshen RA, **Classification and Regression Trees**, Pacific Grove, CA: Wadsworth and Brooks/Cole.

[Buchman *et al.*, 1994]
T.G. Buchman, K.L. Kubos, A.J. Seidler and M.J. Siegforth, **A comparision of Statistical and Connectionist Models for the Prediction of Chronicity in a Surgical Intensive Care Unit**, Critical Care Medicine, 22(5), p750-762: May, 1994.

[Burke, 1996]
H.B. Burke, D.B. Rosen, P.H. Goodman, **Comparing the Prediction Accuracy of Artificial Neural Networks and other Statistical Models for Breast Cancer Survival**, *Advances in Neural Information Processing Systems*, Vol. 8, p1063-1067: 1996.

[Carpenter, 1991]
G.A. Carpenter, S. Grossberg, D.B. Rosen, **Fuzzy ART: Fast Stable Learning and Categorization of Analog Patterns by an Adaptive Resonance System**, *Neural Networks* 4: p759-771: 1991.

[Caruana, Baluja and Mitchell, 1996]

[Charalambous, 1992]
C. Charalambous, **Conjugate Gradient Algorithm for Efficient Training of Artificial Neural Networks**, *IEE Proc.*, Vol.139, No.3, p301-310, 1992.

[Dawant *et al.*, 1993]
B. M. Dawant, S. Uckun, E. J. Manders and D. P. Lindstrom, **The SIMON Project: Model-based Signal Acquisition, Analysis and Interpretation in Intelligent Patient Monitoring**, *IEEE Engineering in Medicine and Biology Magazine*, 12(4), p82-91: 1993.

[De Courcy *et al.*, 1995]
R.H.B. De Courcy-Wheeler, C.D.A. Wolfe, A. Fitgerald, *et al.*, **Use of CRIB Score in Prediction of Neonatal Mortality and Morbidity**, *Arch. Dis. Child Fetal Neonatal*, F32-36, p73: 1995.

[Delogu, in press]
P. Delogu, **Neural Networks for Automatic Diagnosis of Labeled Monodispersed Aerosols in Obstructive Lung Disease**, *Physica Medica*.

[Dempster *et al.*, 1977]
A. Dempster, N. Laird and D. Rubin, **Maximum Likelihood from Incomplete Data via the EM Algorithm**. *Journal of the Royal Statistical Society*, Series, B., 39, 1-38.

[Demuth & Beale, 1997]
H. Demuth, M. Beale, Matlab: Neural Network Toolbox User's Guide, The Math Works Inc.: Natick, MA: January, 1997.

[East *et al.*, 1995]
T.D. East, J. Wallace, A.H. Morris, *etc.*, **Computers in Critical Care**, *Critical Care Nursing Clinics of North America*, Vol. 7, No. 2: June, 1995.

[Ennett and Frize, 1998]
C.M. Ennett and M. Frize, **Investigation into the Strengths and Limitations of Artificial Neural Networks: An Application to an Adult ICU Patient Database**, *AMIA 1998 Annual Fall Symposium*, Orlando, FL. : November, 1998.

[Escobar *et al.*, 1995]
G.J. Escobar, A. Fischer, D.K. Li, *et al.*, **Score for Neonatal Acute Physiology: Validation in Three Kaiser-Permanent Neonatal Intensive Care Units**, *Pediatrics*, p918-922: 1996.

[Famili *et al*, 1997]
A. Famili, Shen Wei-Min, Richard Weber, Evangelos Simoudis, **Data Preprocessing and Intelligent Data Analysis**, *Intelligent Data Analysis*, Vol. 1, No. 1: 01/1997.

[Famili and Turney, 1991]
A. Famili and P. Turney, **Intelligently Helping Human Planner in Industrial Process Planning**, AIEDAM 5(2), p109-124: 1991

[Fayyad *et al.*, 1996]
U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, **From Data Mining to Knowledge Discovery**, in U. Fayyad *et al*, eds., *Advances in Knowledge Discovery and Data Mining*, AAAI/MIT Press, Menlo Park, CA, p1-34: 1996.

[Frize *et al.*, 1995]
M. Frize, F.G. Solven, M. Stevenson, B.G. Nickerson, T. Buskard and K. Taylor, **Computer-Assisted Decision-Support Systems, for Patient Management in a Intensive Care Unit**, *Proceedings of Medinfo '95*, Vancouver, B.C., p1009-1012: July, 1995.

[Frize *et al.*, 1996]
M. Frize, F.G. Solven, M. Stevenson, B.G. Nickerson and H.C.E. McGowan, **Information Technologies Approach and Development for Various Medical Applications**, *Proceedings of the 1996 Canadian Conference on Electrical & Computer Engineering (CCECE'96)*, Calgary, Alberta, p365-368: May, 1996.

[Frize *et al.*, 1998]
Research Proposal in Application for MRC Grand.

[Gray *et al.*, 1992]
J.E. Gray, D.K. Richardson, M.C. McCormick, *et al.*, **Score for Neonatal Acute Physiology (SNAP) and risk of intra-ventricular hemorrhage (IVH)**, *Ped. Res.*, 249A, p31: 1992.

[Hagan & Menhaj, 1994]

M.T. Hagan and M. Menhaj, Training Feedforward Networks with the Marquardt algorithm, IEEE Transactions on Neural Networks, Vol.5, No.6, p989-993: 1994.

[Harrell *et al.*, 1988]
F.E. Harrell, K.L. Lee, B.G. Pollock, **Regression Models in Clinical Studies: Determining Relationships Between Predictors and Response**, *Journal of National Cancer Institute* 80: 1198-1202: 1988

[Henry *et al.*, 1995]
S.B. Henry, K.J. Dolter and C.A. Reilly, Critical Care Information Systems – Essential Infrastructure for Critical Care Nursing Practice, *Critical Care Nursing Clinics of North America*, Vol. 7, No. 2: June, 1995.

[Ho, 1998]
H. Ho, **A Study of the Performance of Artificial Neural Networks to Estimate Outcomes in the Intensive Care Unit**, Undergraduate Senior Thesis, University of Ottawa, Ottawa, ON: 1998.

[Kleinbaum, Kupper, Muller *et al.*, 1998]
D.G. Kleinbaum, L.L. Kupper, K.E. Muller, *et al.*, ***Applied Regression Analysis and Other Multivariable Methods***, Duxbury Press: 1998.

[Knaus *et al.*, 1981]
W. A. Kanus, J. E. Zimmerman, D. P. Wagner, E. A. Draper, D. E. Lawrence, **APACHE-Acute Physiology and Chronic Health Evaluation: A Physiologically-based classification system**, *Critical Care Medicine,* 9, p591-597: 1981.

[Jacobs, 1988]
R.A. Jacobs, **Increased Rates of Convergence Through Learning Rate Adaptation**, *Neural Networks*, Vol.1, No.4, p295-308: 1988.

[John, G.H. *et al.*, 1994]
G.H. John, R. Kohavi and K. Pfleger, **Irrelevant Features and the Subset Selection Problem**, *Proceedings of the 11th IML*, Morgan Kaufmann, 1994.

[Johnston *et al.*, 1994]
M. E. Johnston, K. B. Langton, R. B. Haynes, A. Mathieu, **Effects of Computer-based clinical decision-support Systems on Clinician Performance and Patient Outcome**, *Annals of Internal Medicine*, Vol.120, p135-142: 1994.

[Knaus *et al.*, 1981]
W. A. Kanus, J. E. Zimmerman, D. P. Wagner, E. A. Draper, D. E. Lawrence, **APACHE-Acute Physiology and Chronic Health Evaluation: A Physiologically-based classification system**, *Crit. Care Med*, 9, p591-597:1981.

[Kohavi, 1995]
R. Kohavi, **Wrappers for Performance Enhancement and Oblivious Decision Graphs**, Ph.D. Thesis: 1995.

[Kollias and Anastassiou, 1989]
S. Kollias and D. Anastassiou, **An Adaptive Least Squares Algorithm for the Efficient Training of Artificial Neural Networks**, *IEEE Trans. Circ. Syst.*, Vol.36, No.8, p1092-1101: 1989.

[Lau 1994]
F. Lau, **Development and Validation of a Decision-support System for Cardiovascular Intensive Care**, *Canadian Medical Informatics*, p28-29: 1994.

[Lee, 1996]
H.C. Lee, **Skin Cancer Diagnosis Using Hierarchical Neural Networks And Fuzzy Logic**, Master's thesis, Dept. of Computer Science, University of Missouri-Rolla: 1996.

[Lee, 1998]
S. Lee, **Canadian Neonatal Intensive Care Unit Network 1996-1997**, Center for Health Evaluation Research.

[LeGall *et al.*, 1995]
J. LeGall *et al.*, **Customized Probability Models for Early Severe Sepsis in Adult Intensive Care Patients**, *JAMA*, 273(8), 1995, p644-650:1995.

[Liu and Nocedal, 1992]
D.C. Liu and J. Nocedal, **On the Limited Memory BFGS Method for Large Scale Optimization**, *Math. Prog.*, Vol.45, p503-528, 1989.

[Lowell, Davis, 1994]
W.E. Lowell, G.E. Davis, **Predicting Length of Stay for Psychiatric Diagnosis-related Groups Using Neural Networks**, *JAMIA*, Vol. 1, No. 6: 1994.

[Ma *et* al., 1998]
L. Ma, C. Looney, S. Sukuta, *et al.*, **Tumor Diagnosis Using Backpropagation Neural Network Method**, 1998 Annual Meeting Division of Atomic, Molecular and Optical Physics (DAMOP), Santa Fe, Mexico: May, 1998.

[Marquardt, 1963]
D. Marquardt, **An Algorithm for Least Squares Estimation of Non-linear Parameters**, *J. Soc. Ind. Appl. Math.*, p431-441: 1963.

[McAfee *et al.*, 1998]
A.T. McAfee, B.C. Kaplan, J.K. Hahn, *et al.*, **An Artificial Neural Network in the Diagnosis of Ectopic Pregnancy**, *Global Emergency Medicine Archives*.
http://gema.library.ucsf.edu:8081/Originals/SAEMabs/SA176.html

[Miksch *et al.*, 1996]
S. Miksch, W. Horn, C. Popow, F. Paky, **Utilizing Temporal Data Abstraction for Data Validation and Therapy Planning for Artificially Ventilated Newborn Infants**, *Artificial Intelligence in Medicine*, p543-576:1996

[Neal, 1996]
R. M. Neal, **Bayesian Learning for Neural Networks**, Springer, New York: 1996.

[Pagallo, 1989]
G. Pagallo, **Learning DNF by Decision Trees**, *Proceedings of International Joint Conference on Artificial Intelligence*, Morgan Kaufmann, Palo Alto, CA, Vol. I, p639-644: 1989.

[Pesonen, 1996]
E. Pesonen, M. Eskelinen and M. Juhol, **Comparison of Different Neural Network Algorithms in the Diagnosis of Acute Appendicitis**, *International Journal of biomedical Computing*, 40, p227-233: 1996.

[Pesonen, 1997]
E. Pesonen, **Is Neural Network Better Than Statistical Methods in Diagnosis of Acute Appendicitis**, In C. Pappas, N. Maglaveras, J.-R. Scherrer (eds.): *Medical Informatic Europe'97*, IOS Press, Amsterdam, Netherlands, p377-381: 1997.

[Pesonen, 1998]
E. Pesonen, M. Eskelinen, M. Juhola, **Treatment of Missing Data Values in a Neural Network Based Decision Support System for Acute Abdominal Pain**, *Artificial Intelligence in Medicine*, 13, p139-146: 1998.

[Plate *et al.*, 1997]
T. Plate, P. Band, J. Bert, J. Grace, **A Comparison Between Neural Networks and Other Statistical Techniques for Modeling the Relationship Between Tobacco and Alcohol and Cancer**, *Advances in Neural Information Processing Systems*, MIT Press, Cambridge, MA, Vol. 9, p967-972: 1997.

[Press and Wilson, 1978]
S.J. Press, & S. Wilson (1978). **Choosing between logistic regression and discriminant analysis.** *Journal of the American Statistical Association*, 73, 699-705. This paper sets out to show that logistic regression is better than discriminant analysis and ends up showing that at a qualitative level they are likely to lead to the same conclusions. But it is very useful for clarifying terms.

[Puskorius and Feldkamp, 1991]
G.V. Puskorius and L.A. Feldkamp, **Decoupled Extended Kalman Filter Training of Feedforward Layered Networks**, *Proc. IJCNN*, Vol. I, p771-777: July 1991.

[Richardson and Tarnow-Mordi, 1994]
D. K. Richardson, W. O. Tarnow-Mordi., **Measuring Illness Severity in Newborn Intensive Care**, *J. Intensive Care Medicine*, 9, p20-33: 1994.

[Roberts *et al.*, 1994]
J. Pardey, S. Roberts, L. Tarassenko and D. Siegwart, **A Confidence Measure for Artificial Neural Networks**, In *International Conference Neural Networks and Expert Systems in Medicine and Healthcare*, Plymouth, UK, p23-30: 1994.

[Rogers, 1997]

J. Rogers, *Object-Oriented Neural Networks in C++*, Academic Press, Inc.: 1997.

[Rumelhart , Hinton and Williams, 1986A]
D.E. Rumelhart, G.E. Hinton and R.J. Williams, **Learning Representations by Back-Propagating Errors,** *Nature*, Vol.323, p533-536: 1986.

[Rumelhart, Hinton and Williams, 1986B]
D.E. Rumelhart, G.E. Hinton, and R.J. Williams, **Learning Internal Representations by Error Propagation.** In D.E. Rumelhart and J.L. McClelland, eds., *Parallel Distributed Processing*: Explorations in the Microstruction of Cognition, Vol.1: Foundations, Cambridge, Massachusetts, MIT Press: 1986.

[Rumelhart and McClelland, 1986]
D.E. Rumelhart and J.L.McClelland, *Parallel Distributed Processing*, vol. 1, MIT Press, 1986.

[Sarle, 1996]
Various Authors and W. Sarle Maintainer, **The comp.ai.neural-nets Frequently Asked Questions (FAQ) List**, June 27, 1996 Version. URL: fttp://ftp.sas.com/pub/neural.

[Shanno, 1990]
D.F. Shanno, **Recent Advances in Numerical Techniques for Large-scale optimization**, *Neural Networks for Control*, Miller, Sutton and Werbos, Eds. Cambridge MA: MIT Press: 1990.

[Singhal and Wu, 1989]
S. Singhal and L. Wu, **Training Multilayer Perceptrons with the Extended Kalman Algorithm**, *Advances in Neural Information Processing Systems 1*, D.S. Touretzky, Ed. San Mateo, CA: Morgan Kaufman, p133-140: 1989.

[Specht, 1990]
D.F. Specht, **Probabilistic Neural Networks**, Neural Networks 3: p109-118: 1990.

S-PLUS, v.3.0. Seattle, WA; Statistical Sciences, Inc.: 1991

[Stensmo and Sejnowski, 1995]
M. Stensmo and T.J. Sejnowski, **A Mixed Model System for Medical and Machine Diagnosis**, In: G. Tesauro, D.S. Touretzky and T.K. Leen (Eds.), *Advances in Neural Information Processing Systems*, Vol. 7, MIT Press, Cambridge, MA, USA: 1995.

[Sykacek, Dorffner, *et al.*,1998]
P. Sykacek, G. Dorffner, P. Rappelsberger, J. Zeitlhofer, **Experiences with Bayesian Learning in a Real World Application**, to appear in: Advances in Neural Information Processing Systems, Vol.10, 1998. http://www.ai.univie.ac.at/~georg/papers/nips97.ps.Z

[Tong, Frize and Ennett, 1998]
Y. Tong, M. Frize and C.M. Ennett, **Discussion on the Use of Information Theory to Classify Medical Outcomes**, *AIMA 1998 Annual Fall Symposium*, Orlando, FL. Fs: November, 1998.

[Tong, Frize and Walker, 1999]
Y. Tong, M. Frize and R. Walker, **Estimating Ventilation Using Artificial Neural Networks in Intensive Care Units**, submitted to 1999 Annual Fall meeting of the Biomedical Engineering Society and the 21st Annual International Conference of the IEEE Engineering in Medicine and Biology Society, October 13-16, 1999 Atlanta, GA.

[Tresp, Ahmad and Neuneier, 1994]
V. Tresp, S. Ahmad and R. Neuneier, **Training Neural Networks with Deficient Data**, *Advances in Neural Information Processing Systems*, Vol. 6, J.D. Cowan, G. Tesauro and J. Alspector eds., San Mateo, CA, Morgan Kaufman.

[Tresp, Neuneier and Ahmad, 1996]
V. Tresp, R. Neuneier and S. Ahmad, **Efficient Methods for Dealing with Missing Data in Supervised Learning**, In: G. Tesauro, D.S. Touretzky and T.K. Leen (Eds.), *Advances in Neural Information Processing Systems*, Vol. 7, MIT Press, Cambridge, MA, USA: 1996.

[Trigg, 1997]
H.C.E. Trigg, **An Investigation of Methods to Enhance the Performance of Artificial Neural Networks Used to Estimate ICU Outcomes**, Master's Thesis, Dept. of Electrical Engineering, Univ. of New Brunswick, Fredericton (NB), 1997.

[Vogl et. al., 1988]
T.P. Vogl, J.K. Mangis, A.K. Zigler, W.T. Zink and D.L. Alkon, **Accelerating the Convergence of the Backpropagation Method**, *Bio. Cybern.*, Vol.59, p256-264: Sept. 1988.

[Walker, Frize and Tong, 1999]
R. Walker, M. Frize and Y. Tong, **Data Analysis Using Artificial Neural Networks and Computer-aided Decision-making in the Neonatal Intensive Care Unit**, Paediatric Academy Society 1999 Annual Meeting, San Francisco: May 2, 1999, published in *Paediatr. Res.* 45, p231A: 1999.

[Walker, Frize and Tong, 1999]
R. Walker, M. Frize and Y. Tong, **Data Analysis Using Artificial Neural Networks and Computer-aided Decision-making in the Neonatal Intensive Care Unit**, Canadian Paediatric Society 1999 Annual Meeting, Winnepeg: June 24, 1999, to be published in *Paediatrics and Child Health*. 45, p231A: 1999.

[Weigend *et al.*, 1991]
A. Weigend, D. Rumelhart and B. Huberman, **Generalization by Weight-Elimination with Application to Forecasting**. In *Advances in Neural Information Processing System*, Vol.3, p875-882. Morgan Kaufmann, San Mateo: 1991.

[Werbos, 1988]
P. Werbos, **Backpropagation: Past and Future**. In *Proceedings of the IEEE International Conference on Neural Networks*, p343-353, IEEE Press: 1988.

[Wu *et al.*, 1993]

Y. Wu *et al.*, **Artificial Neural Networks in Mammography: Application to Decision Making in the Diagnosis of Breast Cancer**, *Radiology*, 187, p81-87: 1993.

[Yale, 1997]
K. Yale, **Preparing the Right Data Diet for Training Neural Networks**, *IEEE Spectrum*, p64-66: March, 1997.

[Yu *et al.*, 1982]
V. L. Yu, L. M. Fagan, S. W. Bennett., W. J. Clancey, A. C. Scott, J. F. Hannigan, R. L. Blum, B. G. Buchanan and S. N. Cohen, **Antimicrobial Selection by a Computer: A Blinded Evaluation by Infectious Disease Specialists**, *JAMA*, vol. 242, no. 12, p1279-1282: 1982.

APPENDIX 1: SCORING SYSTEMS

Table 1a: CRIB Score

Factor	Score
Birthweight (g)	
>1350	0
851-1350	1
701-850	4
<=700	7
Gestation (wk)	
>24	0
<=24	1
Congenital Malformation	
None	0
Not acutely life-threatening	1
Acutely life-threatening	3
Maximum base excess in first 12 h (mmol/L)	
>=-70	0
-70 to –99	1
-100 to 140	2
<=150	3
Minimum appropriate Flo2 in first 12 h	
<=0.40	0
0.41-0.60	2
0.61-0.90	3
0.91-1.00	4
Maximum appropriate Flo2 in first 12 h	
<=0.40	0
0.41-0.80	1
0.81-0.90	3
0.91-1.00	5

Table 1b: NTISS Score

Respiratory Subscore	**Cardiovascular Subscore**
Supplemental O2 Surfactant administration Tracheostomy care Tracheostomy placement Continuous Positive Airway Pressure administration Endotracheal intubation Mechanical ventilation Mechanical ventilation with muscle relaxation High frequency ventilation Extra-Corporal Membrane Oxygenation	Indomethacin administration Volume expansion (<=15 cc/kg) Vasopressor administration (one agent) Volume expansion (>15 cc/kg) Vasopressor administration (more than one agent) Pacemaker on standby Cardiopulmonary resuscitation
Drug Therapy Subscore	**Monitoring Subscore**
Antibiotic administration (<=2 agents) Diuretic administration (enteral) Steroid administration (post-natal) Anti-convulsant administration Aminophylline administration Other unscheduled medication Antibiotic administration (>2 agents) Diuretic administration (parenteral) Treatment of metabolic acidosis Potassium binding resin administration	Frequent Vital Signs Cardiorespiratory monitoring Phlebotomy (5-10 blood draws) Thermo-regulated environment Non-invasive O2 monitoring Arterial pressure monitoring Central venous pressure monitoring Urinary catheter Quantitative I&Os Extensive Phlebotomy (>10 blood draws)
Metabolic/Nutrition Subscore	**Transfusion Subscore**
Gavage feeding Intravenous fat emulsion Intravenous amino acid solution Phototherapy Insulin administration Potassium infusion	Intravenous gamma globulin Red Blood Cell transfusion (<=15 cc/kg) Partial volume exchange transfusion Red Blood Cell transfusion (>15 cc/kg) Platelet transfusion White Blood Cell transfusion Double volume exchange transfusion
Procedural Subscore	**Vascular Access Subscore**
Transport of patient Single chest tube in place Minor operation Multiple chest tubes in place Thoracentesis Major operation Pericardiocentesis Pericardial tube in place Dialysis	Peripheral intravenous line Arterial line Central venous line

Table 1c: SNAP-PE Score

Add the followings on the basis of SNAP score.

1. Apar score at 5 min----------apar5 (flat table).
2. SGA (small for the gestational age) not in our data set which is defined as having a birthweight of less than the 3rd percentile for that gestational age. You can calculate by birth weight and gestation age.
3. Birthweight----------bthwgt (flat table).

APPENDIX 2 CONCEPTS IN ASSESSMENT OF DIAGNOSTIC TECHNOLOGIES

Source: Thornbury and Fryback 1992. Thornbury, J. R., and D. G. Fryback. "Technology Assessment – an American View." European Journal of Radiology 14, no. 2 (1992): 147-56.
U.S. National Library of Medicine (NLM)
http://www.nlm.nih.gov, Last updated: 14 May 1998

The relationship between most preventive, therapeutic, and rehabilitative technologies and health outcomes is direct. The relationship between the use of diagnostic and screening technologies and health outcomes is typically indirect; these technologies provide information that may be used to inform providers concerning the use of interventions that may in turn affect health outcomes.

Many tests and other technologies used for diagnosis are also used for screening, and the concepts discussed here for diagnostic technologies pertain as well to screening technologies. A basic difference between screening and diagnosis is that diagnosis is done in symptomatic patients and screening is typically done in asymptomatic patient groups. For a given test used for either screening or diagnosis, this difference has a great effect on the probability that a patient has a disease or other health condition.

The immediate purpose of a diagnostic test is to provide information about the presence (and, less often, the extent) of a disease or other health condition. That is, the diagnostic test should be able to discriminate between patients who have a particular disease and those who do not have the disease (or discriminate among different extents of disease in a given patient).

The technical performance of a diagnostic test depends on a number of factors. Among these are the precision and accuracy of the test, the observer variation in reading the test data, and the relationship between the disease of interest and the cutoff level of the marker or surrogate used in the diagnostic test to determine the presence or absence of that disease. These factors contribute to the ability of a diagnostic test to detect a disease when it is present and to not detect a disease when it is not present.

The marker for a disease or condition is typically defined as a certain cutoff level of a variable such as blood pressure or glucose level. Disease markers have distributions in nondiseased and diseased populations. For most diseases, these distributions overlap, so that a single cutoff level does not clearly separate nondiseased from diseased people. For instance, in the case of the disease of hypertension, a usual marker for the disease is diastolic blood pressure, the cutoff level of which is often set at 95mm Hg. In fact, some people whose diastolic blood pressure is above 95mm will not be hypertensive (false positives, as noted below), and some people with diastolic blood pressure below 95mm will be hypertensive (false negatives, as noted below). Lowering the cutoff to 90mm will decrease the number of false positives, but increase the
number of false negatives.

A diagnostic test can have four basic types of outcomes, as shown in Box 26. A true positive diagnostic test result is one that detects a marker when the disease is present. A true negative test result is one that does not detect the marker when the disease is absent. A false positive test result is one that detects a marker when the disease is absent. A false negative test result is one that does not detect a marker when the disease is present.

Operating characteristics of diagnostic tests and procedures are measures of the technical performance of these technologies. These characteristics are based on the probabilities of the four possible types of outcomes of a diagnostic test. The two most commonly used operating characteristics of diagnostic tests are sensitivity and specificity. Sensitivity measures the ability of a test to detect disease when it is present. Specificity measures the ability of a test to correctly exclude disease in a nondiseased person. One economical way of depicting these operating characteristics for a given diagnostic test is with a receiver operating characteristic (ROC) curve, which

plots the relationship between the true positive ratio (sensitivity) and false positive ratio (1 - specificity) as a function of the cutoff level of a disease (or condition) marker. ROC curves help to demonstrate how raising or lowering the cutoff point for defining a positive test result affects tradeoffs between correctly identifying people with a disease (true positives) and incorrectly labeling a person as positive who does not have the condition (false positives).

Taken alone, sensitivity and specificity do not reveal the probability that a given patient really has a disease if the test is positive, or the probability that a given patient does not have the disease if the test is negative. These probabilities are captured by two other operating characteristics. Predictive value positive is the proportion of those patients with a positive test result who actually have the disease.

Predictive value negative is the proportion of patients with a negative test result who actually do not have the disease. (See Box 27.) Unlike sensitivity and specificity, predictive value positive and predictive value negative are not constant performance characteristics of a diagnostic test; they change with the prevalence of the disease in the population of interest. For example, if the disease is sufficiently rare in the population, even tests with high sensitivity and high specificity can have low predictive value positive, generating more false-positive than false negative results.

Beyond technical performance of diagnostic technologies, the effect of diagnostic technologies on health outcomes or health-related quality of life is less obvious than for other types of technologies. As health care decision makers increasingly demand to know how health care interventions affect health care outcomes, diagnostic technologies will have to demonstrate their efficacy/effectiveness accordingly.

The efficacy (or effectiveness) of a diagnostic technology can be determined along a chain of inquiry that leads from technical capacity of a technology to changes in patient health outcomes to cost effectiveness, as follows.

- Technical capacity. Does the technology perform reliably and deliver accurate information?
- Diagnostic accuracy. Does the technology contribute to making an accurate diagnosis?
- Diagnostic impact. Do the diagnostic results influence use of other diagnostic technologies, e.g., does it replace other diagnostic technologies?
- Therapeutic impact. Do the diagnostic findings influence the selection and delivery of treatment?
- Patient outcome. Does use of the diagnostic technology contribute to improved health of the patient?
- Cost effectiveness. Does use of the diagnostic technology improve the cost effectiveness of health care compared to alternative interventions?

If a diagnostic technology is not efficacious at any step along this chain, then it is not likely to be efficacious at any later step. Efficacy at a given step does not imply efficacy at a later step (Feeny *et al.* 1986; Fineberg *et al.* 1977; Institute of Medicine 1985). Box 28 shows a hierarchy of studies for assessing diagnostic imaging technologies that is consistent with the chain of inquiry noted above.

For diagnostic technologies that are still prototypes or in other early stages of development, there are little data upon which to base answers to questions such as these. Even so, investigators and advocates of diagnostic technologies should be prepared to describe, at least qualitatively, the ways in which the technology might affect diagnostic accuracy, diagnostic impact, therapeutic impact, patient outcomes and cost effectiveness; how these effects might be measured; approximately what levels of performance would be needed to successfully implement the technology; and how further investigations should be conducted to make these determinations.

Table 2a Possible Outcomes of Diagnostic Tests

Test Result	Disease Status	
	Present	Absent
Positive	True positive	False positive
Negative	False negative	True negative

Table 2b Operating Characteristics of Diagnostic Tests

CHARACTERISTIC	FORMULA	DEFINITION

Sensitivity	True positives / True positives + False negatives	Proportion of people with condition who test positive
Specificity	True Negatives / True negatives + False positives	Proportion of people without condition who test negative
Predictive value Positive	True positives / True positives + False positives	Proportion of people with positive test who have condition
Predictive value Negative	True negatives / True negatives + False negatives	Proportion of people with negative test who do not have condition

Table 2c Hierarchical Model of Efficacy for Diagnostic Imaging (Typical Measures of Analysis)

Level 1. Technical efficacy
Resolution of line pairs
Modulation transfer function change
Gray-scale range
Amount of mottle
Sharpness

Level 2. Diagnostic accuracy efficacy
Yield of abnormal or normal diagnoses in a case series
Diagnostic accuracy (% correct diagnoses in case series)
Sensitivity and specificity in a defined clinical problem setting
Measures of area under the ROC curve

Level 3. Diagnostic thinking efficacy
Number (%) of cases in a series in which image judged "helpful" to making the diagnosis
Entropy change in differential diagnosis probability distribution
Difference in clinicians' subjectively estimated diagnosis probabilities pre- to post-test information
Empirical subjective log-likelihood ratio for test positive and negative in a case series

Level 4. Therapeutic efficacy
Number (%) of times image judged helpful in planning management of patient in a case series
% of times medical procedure avoided due to image information
Number (%) of times therapy planned before imaging changed after imaging information obtained (retrospectively inferred from clinical records)
Number (%) of times clinicians' prospectively stated therapeutic choices changed after test information obtained

Level 5. Patient outcome efficacy
% of patients improved with test compared with/without test
Morbidity (or procedures) avoided after having image information
Change in quality-adjusted life expectancy
Expected value of test information in quality-adjusted life years (QALYs)
Cost per QALY saved with imaging information
Patient utility assessment; e.g., Markov modeling; time trade-off

Level 6. Societal efficacy
Benefit-cost analysis from societal viewpoint
Cost-effectiveness analysis from societal viewpoint

APPENDIX 3: PUBLICATIONS

[Tong, Frize and Walker, 1999]

Yanling Tong, Monique Frize and Robin Walker, **Estimating Ventilation Using Artificial Neural Networks in Intensive Care Units**, submitted to 1999 Annual Fall meeting of the Biomedical Engineering Society and the 21st Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Atlanta, GA.: October 13-16, 1999. Student Competition Regional Finalist, Award of Merit.

[Tong, Frize and Ennett, 1998]

Yanling Tong, Monique Frize and Colleen M. Ennett, **Discussion on the Use of Information Theory to Classify Medical Outcomes**, *AIMA 1998 Annual Fall Symposium*, Orlando, FL. Fs: November, 1998.

[Walker, Frize and Tong, 1999]

Robin Walker, Monique Frize and Yanling Tong, **Data Analysis Using Artificial Neural Networks and Computer-aided Decision-making in the Neonatal Intensive Care Unit**, Paediatric Academy Society 1999 Annual Meeting, San Francisco: May 2, 1999, published in *Paediatr. Res.* 45, p231A: 1999.

[Walker, Frize and Tong, 1999]

Robin Walker, Monique Frize and Yanling Tong, **Data Analysis Using Artificial Neural Networks and Computer-aided Decision-making in the Neonatal Intensive Care Unit**, Canadian Paediatric Society 1999 Annual Meeting, Winnepeg: June 24, 1999, to be published in *Paediatrics and Child Health*. 45, p231A: 1999.

ESTIMATING VENTILATION USING ARTIFICIAL NEURAL NETWORKS IN INTENSIVE CARE UNITS

Yanling Tong*, Monique Frize*# and Robin Walker@
*S.I.T.E., Univ. of Ottawa, Ottawa, Canada
*# System & Comp. Eng., Carleton Univ., Ottawa, Canada
@Division of Neonatology, Children's Hospital of East Ontario, Ottawa, Canada

Abstract—Few medical researchers have achieved correct classification Rates (CCR) with Artificial Neural Networks (ANNs). Trigg *et al.* [1] obtained good results using an Adult Intensive Care Unit (AICU) database. This paper extends Trigg's technique – ANNs with weight-elimination, to estimate ventilation with a Neonatal Intensive Care Unit (NICU) database. Encouraging results were obtained in terms of Correct Classification Rate (CCR) and Average Squared Error (ASE) and the adult model could be successfully applied to neonatal patients.

Index Terms—ANN, ventilation, adult & neonatal ICU.

BACKGROUND

The NICU is a fast tempo environment and an expensive resource in today's hospitals. Therefore, it is important to make medical decisions accurately, swiftly, and economically to identify patients at high risk. This requirement has led to the interest in determining the relative importance of various factors contributing to resource utilization. The Score for Neonatal Acute Physiology (SNAP) II, a new version of SNAP, is starting to be widely used in NICUs, particularly in Canada. This work assesses the performance of ANNs in estimating ventilation from admission day SNAP II variables for NICU patients.

METHODOLOGY

Between January 8, 1996 and October 31, 1997, there were about 20,000 admissions of babies in 17 Canadian NICUs. The database used in this research consists of information on 6,905 NICU infants from 5 of these hospitals. Of the 6,905 records, 5,584 have missing information in some fields. These incomplete records were not included in these results as no simple, accurate method to deal with them in ANNs has yet been found.

A feed-forward backpropagation double-layered ANN model was built. It had six inputs (low blood pressure, low temperature, low serum pH, presence of seizures, urine output and PaO2/FiO2 ratio) and one output (ventilation). The model was tested for ventilation <8 hrs. Of the 1,321 complete records, 441 were set aside for testing, leaving 880 records for training, according to the pattern "train, train, test, train, train, test, ...". Each network was trained three times from different small random starting weights with and without weight-elimination (to control overfitting). The one with the best performance on the training data was selected as "the model". The values of CCR and ASE were recorded for comparison.

RESULTS

From the results in Table 1, the ANNs estimate consistently better than the Constant Predictor (CP, a simple statistical benchmark that classifies all patterns as belonging to the class with the highest training set a priori probability) in terms of CCR and ASE. The performance of the model without weight-elimination is further improved by controlling overfitting with weight-elimination (WE).

The results of AICU obtained by Trigg *et al.* were for the same output, but with 6 different variables (heart rate, respiratory rate, fraction of inspired Oxygen, partial pressure of Oxygen in the blood, arterial pH, and glasgow coma score) using the database from the Doctor Everett Chalmers Hospital.

Table 1 shows the CCR, ASE, number of epochs and CP for both adult & neonatal database.

Table 1: Predicting Ventilation (<8 hrs) in NICU & AICU by ANNs

	# of Variables	MAX CCR	ASE of MAX CCR	# of Epochs	CP
AICU [1] without WE	6	90.5%	0.32	130	71%
NICU with WE	6	92.7%	0.13	540	61%
NICU without WE		91%	0.2	650	

CONCLUSION

The adult model could be successfully applied to neonates with the current database used. The results also show that ANNs can be useful tools in predicting ventilation based on the admission day SNAP II variables. It improves the performance significantly by using only a small number of input variables.

Future work will consist of investigating effective methods to make full use of the information in the incomplete records. We also intend to extend this approach to the estimation of other NICU outcomes of interest.

REFERENCES

[1] H. Trigg, M. Stevenson and M. Frize (1997), "Estimating Ventilation Requirements in Critical Care Medicine," *Proceedings of the World Congress on Biomedical Engineering and Physics*. Nice, France, September 14-19.

Discussion on the Use of Information Theory to Classify Medical Outcomes

Yanling Tong*, Dr. Monique Frize*# and Colleen M. Ennett*
*U. of Ottawa, S.I.T.E., 161 Louis Pasteur, Ottawa, ON K1N 6N5
#Carleton U., Systems & Comp. Eng., 1125 Colonel By Drive, Ottawa, ON K1N 5B5

As the principles behind information theory (IT) are analogous to clinical thinking, IT appears to be a complementary approach to artificial neural networks (ANNs). This paper addresses the application of IT to classification analysis, to which this technique belongs.

BACKGROUND

Exploring various analytical approaches to process data from medical databases is meaningful before deciding on the tool that will be the most useful, accurate, and relevant for practitioners. For example, assigning a new patient to a particular outcome class is a classification analysis problem commonly described as Apattern recognition,@ Adiscriminant analysis@ and Asupervised learning.@ In this paper IT is adopted to solve this kind of question.

METHODOLOGY

Transinformation (mutual information) refers to the information transmitted through a predictive model, and is a measure of the reduction in uncertainty attributed to using a particular model to classify each vector. Therefore, information quantities like entropy, relative entropy, and mutual information can be used to preprocess information before ANN analysis.

Two projects were carried out on an Adult Intensive Care Unit (AICU) database (1,491 records) and an Neonatal Intensive Care Unit (NICU) database (291 records), respectively [1,2].

For each project, an IT model investigated the relationship between each input of the database and the desired output (in this case, duration of artificial ventilation). By classifying the input variables into several groups, the probability distribution of each group was assessed. Entropy, conditional entropy, joint entropy and mutual information were calculated using these classes as the model inputs. Conditional entropy was the main model parameter. When admitting new patients, assignment to a certain group occurred based on this input value. The output class was predicted by means of model parameters. A large value of mutual information indicates that the input and output are closely interrelated. Zero mutual information means that the input and output are independent, that is, no useful information can be extracted from the model.

DISCUSSION

A limitation of the above method is its representation of the dependent relation between only one input and one output. In practice, all the inputs should be considered to estimate the output. Subsequent to this analysis, a well-established, layered ANN backpropagation algorithm was used in conjunction with IT to evaluate the ANN=s performance.

Another important issue is the size of the database. Size not only influences the results of IT analysis, but it also limits the efficiency and convergence speed of the ANN. The AICU database achieved a 6.5 to 16.75% improvement in correct classification rate comparing to the results to a constant predictor. In the case of the NICU database, the mutual information was zero in most cases indicating the output of the model was independent of the inputs selected for the study.

CONCLUSION

Although the IT model cannot predict individual outcomes, it could be a useful tool to indicate under which circumstances the output happens most frequently. This information could improve the time management of the expensive and limited resources.

Future work will consist of repeating these studies with a much larger database acquired from the Children=s Hospital Of Eastern Ontario (CHEO) in Ottawa, ON.

References

1. Tong Y. Modelling, simulation and performance evaluation. Term Project, University of Ottawa, Ottawa, ON: 1997.
2. Ennett C. Estimating duration of mechanical ventilation in the ICU using artificial neural networks. Term Project, University of Ottawa, Ottawa, ON: 1997.

DATA ANALYSIS USING ARTIFICIAL NEURAL NETWORKS AND COMPUTER-AIDED DECISION-MAKING IN THE NEONATAL INTENSIVE CARE UNIT

C. Robin Walker, MB, ChB, FRCPC, Monique Frize, PhD, PEng, OC, Yanling Tong, MSc

Department of Paediatrics, Children's Hospital of Eastern Ontario, University of Ottawa, Ottawa, ON, Canada
Department of Systems and Computer Engineering, Carleton University, Ottawa, ON, Canada

Introduction

Illness severity measurements such as the neonatal scores CRIB, SNAP and SNAP-II allow prediction of mortality and have been shown to be useful in assessing the probability of other outcomes, including a variety of morbidities and resource utilisation. However, such scores can relate only a limited number of variables to outcome -- indeed, limiting the number of input variables is often a deliberate intent of a score=s originator to make it easier to use. The limited number of input variables may limit the possible outcomes that can be predicted by the score. Moreover, once validated, the score is static and weighting of the variables does not change even if clinical care changes in the Neonatal Intensive Care Unit (NICU) affect the importance of one or more components of the score. Finally, scoring systems have not been shown to be useful to clinicians in managing patient care except as aids to accurate prognosis -- they do not for example provide the clinician with information which might change the diagnosis or management of the patient. We present in this paper preliminary information on our research using two artificial intelligence (AI) applications -- Artificial Neural Networks (ANNs) and Computer-Aided Decision-Making using a Case-Based Reasoner System (CBR).

Artificial Neural Networks

Artificial Neural Networks offer an alternative method for predicting a broad variety of outcomes from large databases. The ANN mimics the human learning and over a large number of >runs= with a database >learns= which variables relate to which outcome. There is essentially no limit to the number of variables that the ANN can study to construct an algorithm predictive of an outcome. Unlike most statistically derived illness severity scores, the ANN can also construct a different algorithm for each different outcome and the algorithms will reflect the data in the database used, so can be specific to local, regional or national care patterns. Moreover, as new data is added to a database, the ANN can be used to reassess the database and update its algorithm/s to reflect changes in importance of input variables.

ANNs are not without disadvantages. They require very large numbers of patients to predict rare outcomes. It is also possible for errors to appear due to >over-fitting= although correction for this can be built into the system. And there is little data yet to compare the accuracy of ANNs with statistical methods. We have used ANNs to analyse for probability of mortality and other relatively common outcomes from an NICU database of 1100 patients collected over a 22-month period from two tertiary NICUs in our centre (one inborn NICU with a population predominantly of Extremely Low Birth Weight and Low Birth Weight babies, one referral NICU with a population of mainly surgical and cardiac cases.) We have analysed for less common outcomes from a larger database which incorporates our centre and data from four other large NICUs. The databases both contain time-varying data entered on days 1, 3 and 15 of each patient=s admission in NICU.

ANN Methodology

Between January 8, 1996 and October 31, 1997, there were about 20,000 admissions of babies in 17 Canadian NICUs (which comprise the Canadian Neonatal Network or CNN.) The database used in this research consists of information on 6,905 NICU infants from 5 of these hospitals. Of the 6,905 records, 5,584 have missing information in some fields. These incomplete records were not included in these results at this time, as no simple, accurate method to deal with them in ANNs has yet been identified.

A feed-forward back-propagation double-layered ANN model was built. It had six inputs (low blood pressure, low temperature, low serum pH, presence of seizures, urine output and PaO2/FiO2 ratio) and one output (ventilation). The model was tested for ventilation <8 hrs. Of the 1,321 complete records, 441 were set aside for testing, leaving 880 records for training, according to the pattern 'train, train, test, train, train, test=, *etc.* Each network was trained three times from different small random starting weights without and with weight-elimination (to control over-fitting). The run with the best performance on the training data was selected as 'the model'. The values of Correct Classification Rates (CCR) and Average Squared Error (ASE) were recorded.

ANN Results

From the results in Table 1, the ANNs estimate consistently better than the Constant Predictor (CP, a simple statistical benchmark that classifies all patterns as belonging to the class with the highest training set a priori probability) in terms of CCR and ASE. The performance of the model without weight-elimination is further improved by controlling overfitting with weight-elimination (WE).

The results of AICU (Adult Intensive Care Unit) obtained by Trigg *et al.* were for the same output, but with 6 different variables (heart rate, respiratory rate, fraction of inspired Oxygen, partial pressure of oxygen in the blood, arterial pH, and Glasgow Coma Score) using a database from the Doctor Everett Chalmers Hospital in Fredericton NB.

Table 1 shows the CCR, ASE, number of epochs and CP for both adult & neonatal database.

Table 1: Predicting Ventilation (<8 hrs) in NICU & AICU by ANNs

	# of Variables	MAX CCR	ASE of MAX CCR	# of Epochs	CP
AICU [1] without WE	6	90.5%	0.32	130	71%
NICU with WE	6	92.7%	0.13	540	61%
NICU without WE		91%	0.2	650	

Decision Support System (Case Based Reasoner)

Although ANNs offer the possibility of a sophisticated outcome-predictor system, in the clinical setting this is of value only in providing the physician with a clearer prognosis and in the assessment of resource utilisation. Such predictions do not provide, at this early stage, the physician information which supports diagnosis or management decisions.

Physicians commonly use experience gained in the management of cases they believe to be similar to a current case to develop a differential diagnosis and to determine which direction their management of the case should take. We hypothesise that a system which accurately provides diagnostic, physiologic and therapeutic information on past cases >matched= to a current case would be of value in supporting the physician=s decisions in the clinical setting.

Such systems can be designed using the Case-Based Reasoner methodology. These systems use a complex series of questions set in the form of >if x . . . then y= or >if not x . . . then z=, *etc.*, to compare a current case with cases from a database. The system can be set up to use weighting of variables according to physician opinion or, for a more scientific match, weights can be assessed through extensive statistical analyses of the data. Any number of variables can be assessed in a database but the system characteristically will determine which and how many are necessary for an accurate match. The physician can then be presented with a pre-specified number of cases matched to the current case and

can review all the information in the database on each such patient to support diagnostic and management decisions. One advantage of the system to the physician is that these patients are drawn from the database presented to the system, i.e. if this is a local database, then the cases will be matched on the basis of data reflecting local practice.

Compared with a physician=s own recollection of similar past cases, we believe the CBR will be more accurate in >matching= cases and able to recall many more cases than a physician would be likely to remember. Thus the information presented is likely to be more accurate and comprehensive than the physician=s own recollection in providing information from which past >experience= can be applied. However, our hypothesis that such a system will be of clinical value will require further testing through a randomised clinical trial once the software is fully developed.

We have developed a prototype decision-support software program for the NICU based on data from two NICUs at one centre from the network described above (the CNN). The software matches a newly entered patient to the closest matching patients in the database using several weighted parameters from the database. (At this time, weighting for this demonstration study has been assigned solely on the basis of the principal author=s opinion.) The matching patients= full data is displayed and thus available to the NICU physician as a potential aid to his/her decision-making on the newly entered patient.

We show below several screens from the current version of our software (IDEAs for NICUs.) Our software is at an early stage of development and we will be refining it further so as to provide clearer information on-screen and to make it more user- (specifically physician-) friendly.

Figure 1

Figure 2

Figure 3

Figure 4

Conclusion

Although AI applications in the NICU are at a relatively early stage of development, it is evident that they offer at least the possibility of more accurate outcome prediction for a very broad range of outcomes and support for clinical decision-making. ANNs can learn from large databases to predict many outcomes in a way that is not easy or possible with conventional statistical methods. The algorithms are individualised to the database used (e.g., local, regional or national) and to each outcome analysed. They are also capable of updating algorithms as needed to reflect changes in patients or care. CBR systems offer the potential of providing the physician with information on matched patients from the physician=s own centre or region, again according to the database used. The systems however need to be refined to be easily used by physicians. Once this is achieved, the clinical utility of such systems will need to be demonstrated through further research.

Funding

National Science and Engineering Council of Canada (NSERC)
Medical Research Council of Canada (MRC)
CHEO? ($5000. Plus time spent on data colletcion)