uOttawa

L'Université canadienne
Canada's university

FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTOCTORALES

uOttawa

L'Université canadienne
Canada's university

FACULTY OF GRADUATE AND POSDOCTORAL STUDIES

Daphné Townsend

AUTEUR DE LA THÈSE / AUTHOR OF THESIS

M.A.Sc. (Electrical Engineering)

GRADE / DEGREE

School of Information Technology and Engineering

FACULTÉ, ÉCOLE, DÉPARTEMENT / FACULTY, SCHOOL, DEPARTMENT

Clinical Trial of Estimated Risk Stratification Prediction Tool

TITRE DE LA THÈSE / TITLE OF THESIS

Michèle Fortier

DIRECTEUR (DIRECTRICE) DE LA THÈSE / THESIS SUPERVISOR

CO-DIRECTEUR (CO-DIRECTRICE) DE LA THÈSE / THESIS CO-SUPERVISOR

EXAMINATEURS (EXAMINATRICES) DE LA THÈSE / THESIS EXAMINERS

Jane Irvine

Luc Pelletier

Elizabeth Kristjansson

Robert Reid

Gary W. Slater

Le Doyen de la Faculté des études supérieures et postdoctorales / Dean of the Faculty of Graduate and Postdoctoral Studies

CLINICAL TRIAL OF ESTIMATED RISK STRATIFICATION PREDICTION TOOL

Submitted by

Daphné Townsend B.A.Sc

Thesis submitted to the
Faculty of Graduate and Postdoctoral Studies
In partial fulfillment of the requirements
For the degree of Master of Applied Science in Electrical Engineering

School of Information Technology and Engineering
Faculty of Engineering
University of Ottawa

Your file Votre référence
ISBN: 978-0-494-49285-7
Our file Notre référence
ISBN: 978-0-494-49285-7

Abstract

This work presents doctors with a model of the estimated degree of risk of rare and important neonatal outcomes to aid in better decisions and improved allocation of equipment and resources. An extensive list of admission day parameters is reduced to minimum variable sets to create models for outcomes that are relevant to decision-making in the neonatal intensive care unit. Models are applied to a special collection of cases and compared to neonatologists' risk estimates. A comparative analysis of physician's predictions and the models' discrimination abilities highlights areas of success and areas that can be improved for future trials. Doctors responded positively to the prediction interface concept and to the estimated risk stratification models. Physicians' strengths identified outcomes that could benefit from increased sensitivity.

A substantial effort was made to conduct the usability and performance evaluations within the ethical standards that are especially important for engineering healthcare management applications.

Acknowledgements

I would like to thank my supervisor Dr. Monique Frize for her constant encouragement in my projects, and for introducing me to numerous contacts who were most helpful in the completion of this thesis and exposed me to different aspects of biomedical engineering.

I would like to thank my parents for their continued support and encouragement over the course of my undergraduate and graduate studies. Further, I would like to thank current and past members of the Medical Information-technologies Research Group for their invaluable help. Without their effort the result of my thesis would have been far less satisfactory.

Last but not least I would like to express my gratitude to Nathan for being a constant source of motivation over the past two years, for putting things into perspective when my inspiration ran low and for allowing me to fully concentrate on my thesis during the last writing stages.

Table of Contents

List of Figures

List of Tables

Nomenclature

ANN Artificial Neural Network

BPD Bronchopulmonary Dysplasia

CBR Case-Based Reasoner

CCR Correct Classification Rate

CDS Clinical Decision Support

CHEO Children's Hospital of Eastern Ontario

CNN Canadian Neonatal Network™

CP Constant Predictor

CRIB Clinical RIsk for Babies score

DOV Duration of Ventilation

EPIC Evidence-based Practice and Change

IVH Intraventricular Hemorrhage

k-NN k-Nearest Neighbour

LOS Length Of Stay in the NICU

MSE Mean Squared Error

MIRG Medical Information Technologies Research Group

NICU Neonatal Intensive Care Unit

NEC Necrotizing Enterocolitis

NTISS Neonatal Therapeutic Intervention Scoring System

ROC Receiver Operating Characteristic

SNAP Score for Neonatal Acute Physiology

TAM Technology Acceptance Model

CHAPTER 1 INTRODUCTION

1.1 Motivation

1.1.1 Neonatal Healthcare Perspective

The neonatal intensive care unit (**NICU**) is a complicated, fast paced environment where decisions have a critical impact on patients. Due to advances in therapeutic interventions and medical knowledge, clinical staff has more information sources at their disposal. Additionally, bedside devices are now connected to servers which collect data at far greater speed than clinicians can currently use. Clinical decision support (**CDS**) systems can be designed for use in the NICU and are currently "advocated as a means of reducing medical errors and improving decision making in health care" [Catley & Frize 2003]. The outcome estimation models developed in this thesis will be compared to physician's estimated risk for a range of important outcomes, and physician's attitudes and comments will be collected. Once CDS systems are refined, they could be interconnected as part of an internet component based health care system integrating electronic patient records and other computerized systems. Along with computing advances, these systems will allow the gap between medical knowledge and healthcare informatics to be bridged.

The diffusion of information technology has been slower in healthcare than in other industries, and studies on hospital and health information systems have documented many barriers (scepticism about usefulness, resistance to changes in protocol, concerns about time pressures and greater administrative workload) [Safran 2003]. It is recognized that a lack of physician consultation during the development process and poor personnel training may be responsible for clinical implementation resistance. Proper usability evaluations are critical for clinical acceptance of new practices or procedures [Ewing et al. 2003], [Kaplan 2001].

1.1.2 Engineering Background

Engineering progress in medical instrumentation, digital imaging and real-time data collection has led to increasing amounts of information being available for analysis.

Presently, engineers are in a position to express the available information in an intelligent and efficient way so that it assists, rather than encumbers, physicians and their patients. A significant amount of knowledge also resides in the wisdom of experienced doctors. It is known that medical experts intuitively use the memory of experiences and individual observations, and compare similar ones to a current case to support a decision of treatment [Frize & Walker 2000]. The challenge of CDS systems is to combine sources of information and extract knowledge to improve certain aspects of healthcare delivery.

Expert systems can be created by combining artificial intelligence techniques with large existing medical databases to select and display cases similar to the one being investigated, to provide knowledge about possible outcomes or to present all relevant information in a comprehensive and helpful way. Engineering know-how can be applied to databases to improve healthcare distribution and management with the design of tools that aid in diagnosis. These tools can be extended to identify new risk factors of illness, to prevent adverse events and to aid treatment with real-time decision support features.

The Medical Information Technologies Research Group (**MIRG**[1]) has been applying engineering techniques of artificial intelligence since the early 1990's to large medical databases of adult and neonatal populations. MIRG has been successful at creating clinical decision support algorithms with neural network learning, and is bringing them one step closer to clinical implementation with a study of clinicians' judgement and usability evaluations in this thesis.

1.1.3 Multidisciplinary Perspective

Cognitive engineers are investigating the relationship between computerized applications and NICU users with respect to the decision process, motivated by the high degree of complexity and uncertainty of decisions in tertiary care centers [Alberdi et al. 2000]. MIRG is also investigating how decision support tools can be used by parents facing difficult decisions about the healthcare options for their newborn with its PArent Decision Support software [Frize et al. 2005]. Healthcare information systems should be

[1] MIRG Principal Investigators: Dr. Monique Frize, P.Eng (University of Ottawa and Carleton University, Ottawa, ON), Dr. C.R. Walker, MBBS, FRCPC (IWK Health Center, Halifax, NS) and Dr. E. Bariciak MD FRCPC (Children's Hospital of Eastern Ontario, Ottawa, ON).

used to facilitate effective communication and collaboration between healthcare providers, the patients and their families [Safran 2003]. A CDS system that improves communication between doctors and parents, and explicitly addresses their needs can increase the satisfaction parents feel with the care being provided to their baby [Safran 2003]. Many usability studies cite the importance of collaborative teams, and shared decision making is now accepted as an ethical imperative in the NICU [Whitney 2003].

Clinical decision support systems also have the potential to improve the quality of medical and nursing care provided to NICU patients [Alberdi et al. 2000], but they expose some ethical issues relating to design, use and purpose. CDS systems should not be designed as a threat to the physician's autonomy, but as a complement to their decision-making strengths. They must be designed well technically, medically and ethically because errors can have grave consequences. Engineers have a responsibility to ensure correct use of these systems by providing training, especially for younger clinicians who have less experience and have seen fewer of the rare outcomes studied. An effort must also be made to prevent the spread of misinformation regarding the system. For example a clinician must not rely solely on the CDS system for decisions as it is designed to provide additional information. It is imperative that CDS systems do not suppress the doctor's decision. Further, having a highly reliable and available CDS system aimed for doctors does not justify an under-qualified person using it to make a doctor's decisions. Most importantly, a CDS system must increase the standard of care by helping patients, without increasing the risk of doctor error.

Certain ethical concerns such as data privacy and security with CDS systems and electronic patient records will be partly overseen by regulations [Courtright et al. 2001]. All aspects of design, implementation and post-implementation must respect that a "fundamental principle of both legal and ethical decision-making is that the best interest of the child are paramount" [Larcher & Hird 2002]. When ethical issues are properly handled, the CDS field can expand and help improve the quality of healthcare provided [Berner 1998].

1.2 Problem Statement

Currently, very few models using either statistical or artificial intelligence techniques achieve a sensitivity and specificity high enough to merit incorporation into clinical practice. This work will present doctors with a more complete model of rare and important outcomes using artificial neural networks learning. Developed models will be applied to a special collection of cases and presented to neonatologists for their assessments. A comparative analysis of physician and prediction models' discrimination abilities highlights areas of success and areas that require improvement prior to clinical implementation of the CDS system.

A substantial effort will be made to conduct the usability and performance evaluations ethically, which is especially important when engineering healthcare applications. This study observes the discrimination ability of physicians in comparison to models of the estimated risk stratification. Physicians' attitudes towards the CDS systems in order to obtain information for refinement of future systems.

1.3 Thesis Objectives

The first goal of this thesis is to determine the effectiveness of outcome-independent imputation of missing values in discharged patient files used in creating prediction models. This will be achieved by determining a mortality model, and comparing it to the model created with outcome specific imputations. Artificial neural networks are used to create models of mortality, length of NICU stay above 28 days, duration of ventilation above 7 days, bronchopulmonary dysplasia (**BPD**), severe intraventricular hemorrhage (**IVH**) and necrotizing enterocolitis (**NEC**) using this database.

The second goal of this thesis is to compare the discrimination ability of physicians' prediction of outcomes in the NICU using a survey with neonatologists. Models provide estimated risk stratification (physician requirement) and use only data available within 12 hours of admission (medical protocol requirement and database constraint). Collecting the attitudes of physicians on this technology provides important new information and will help steer the development of MIRG's real-time CDS system.

The study is conducted in five parts with the following objectives:

1) Confirm the ability of a case-based reasoner making use of a k-nearest neighbour **(k-NN)** algorithm with uniform weights to impute missing values into a dataset by comparing imputed values to known values and estimating the imputation error.

2) Compare the effect of weighted versus uniform imputation on the classification performance and relative variable importance in mortality models.

3) Determine a minimal variable set for each outcome considered using a single database.

4) Use neural network learning with the maximum likelihood estimation to estimate the risk stratification for six important clinical outcomes and classify the risk into three categories: low (0% - 20%), moderate (21% - 74%) and high (75% - 100%).

5) Compare clinicians' judgments to the CDS system in a clinical environment and analyze results to identify areas of success and areas that need improvement. Validate the case data and prediction display by means of a questionnaire to assess physicians' attitudes towards the prototype.

1.4 Thesis Outline

This thesis is divided into the following seven chapters:

Chapter 1: Introduction has presented the motivation, problem statement and objectives.

Chapter 2: Neonatal Medical Environment covers the relevant information pertaining to neonatal medical databases, clinical decision support and outcome prediction for groups and individuals in the NICU.

Chapter 3: Literature Review and Relevant Work presents the necessary mathematical background for the completion of this thesis. Presented are classification-based artificial neural networks and the maximum likelihood cost function and learning, weight update and performance parameters. Other topics such as configurable neural network parameters, data pre-processing, and stopping criteria are introduced.

Chapter 4: Critical Analysis presents a critical analysis of the original imputation of the neonatal database and outcome prediction models that were derived from it. Next is a critical analysis of the pilot survey with physicians and discussion of possible changes.

Chapter 5: Methodology details the major steps of the variable imputation, the determination of the minimal dataset for each outcome. The final sections of this chapter detail the application of the estimated risk stratification models to the survey database and the design of the user interface and questionnaire for the study.

Chapter 6: Results and Discussion presents experimental results and observations obtained in the prediction test and the survey with neonatologists.

Chapter 7: Conclusions summarizes the results and contributions to knowledge. Future work is also suggested.

CHAPTER 2 NEONATAL MEDICAL ENVIRONMENT

2.1 Neonatal Medicine

When death occurs before the age of 28 days, it is defined as neonatal mortality, whereas stillbirth or death in the first 7 days is defined as perinatal mortality. "Premature births [births occurring before 37 weeks of gestation] are associated with many problems of major clinical significance" like neurodevelopmental problems, chronic respiratory problems, infections and ophthalmologic problems [Khalil et al. 1995], [McLaughlin et al. 1999].

Premature babies account for approximately 70% of the populations in Canada's NICUs [CNN 2004]. In Canada, as in many other countries, the preterm birth rate has been increasing over the past 20 years, and reached 7.1% in 1996 [McLaughlin et al. 1999]. These babies account for 75-85% of all perinatal mortality in Canada, the likelihood of which increases as birth weight (or gestational age) decreases [McLaughlin et al. 1999]. The increase in premature babies is largely attributed to the rise in in-vitro fertilization which often results in multiple births, obstetrical interventions, higher registration of extremely early-gestation births, ultra-sound based estimates of gestational age and an increase in the sophistication of technology (high-frequency ventilators, inhaled nitric oxide, surfactants...) [Joseph et al. 1998]. MIRG is one of the research groups working towards identifying risk factors for preterm birth such as the incidence of maternal smoking, genital tract infection and pre-eclampsia, which are crucial for improving pregnancy outcomes [Catley et al. 2005].

The limit of viability is currently around 25 weeks of gestation (meaning a birth weight of 775g, with a standard deviation of 150 grams [Kramer et al. 2001]) [Van Riper 2001]. Of 1016 infants studied between 1993 and 1999 with a gestational age of 24 weeks or less or a birth weight of 750 g or less, only 246 (24%) survived. Of these, "30% had cerebral palsy, 5% had hearing impairment, and 2% were blind" [Shankaran et al. 2004]. This serves to illustrate the uncertainty that surrounds the mortality and morbidity outcomes about this category of babies because only one quarter of babies survive, and 30% of them do so with a high degree of disability or morbidity.

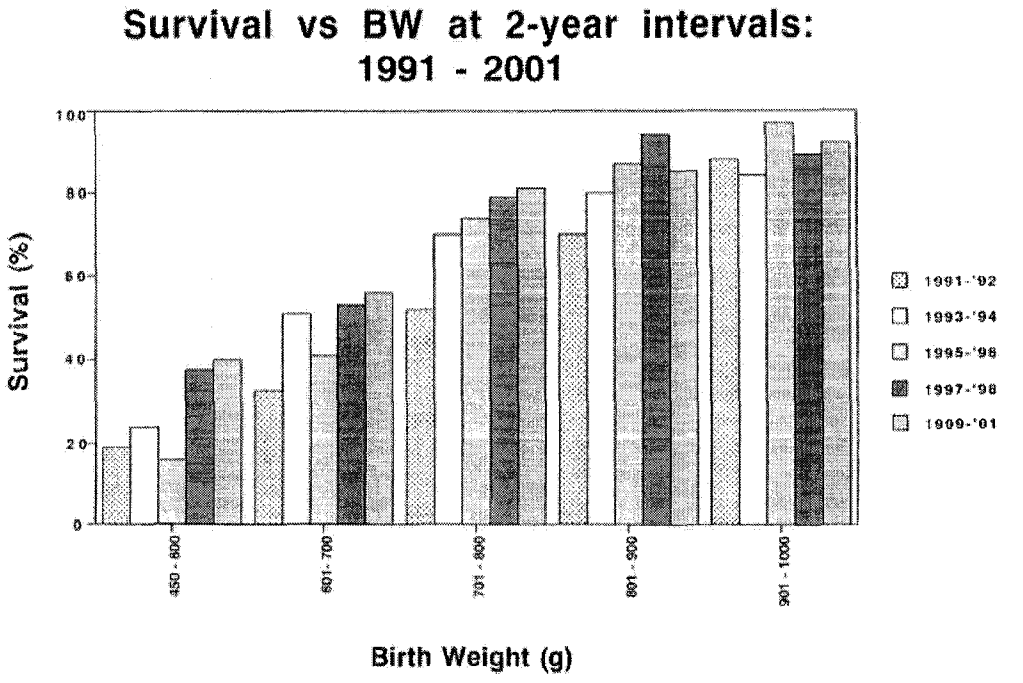

Figure 2.1. Chance of survival for the smallest infants [Meadow et al. 2004]

Figure 2.1 illustrates the improving survival rates for babies in the lowest birth weight categories over the course of a decade. Morbidity percentages in the survivors however have not been decreasing, and a trend towards increasingly aggressive intensive care is cited as the reason for the increased survivors. The three most commonly occurring complications in the newborn population are:

Bronchopulmonary Dysplasia (BPD): a chronic lung disorder characterized by inflammation and scarring of the lungs most common in babies born prematurely and receiving prolonged mechanical ventilation [Chien et al. 2002].

Severe Intraventricular Hemorrhage (IVH): bleeding in the brain, which causes an increase in the intracranial pressure. In mild cases, (grades 1 and 2) the blood is reabsorbed by the brain and in some severe cases (grades 3 and 4), it can lead to brain damage. Severe IVHs are strongly associated with long-term disability. The two primary factors contributing to IVH are poor cerebral auto regulation and abrupt changes in cerebral blood flow and pressure. The first factor is likely due to prematurity and the second can be due to seizures, hypoxia and a competition between mechanical ventilation

and spontaneous breathing events [Annibale & Hill 2006]. "With increasing survival of preterm infants, reducing the incidence of morbidity (e.g., chronic lung disease and severe intraventricular hemorrhage) is a priority in the neonatal intensive care unit" [Chien et al. 2002]. BPD and IVHs are most relevant to infants of ≤32 weeks gestational age [Chien et al. 2002].

Necrotizing Enterocolitis (NEC): a serious intestinal illness affecting primarily babies under 1500 grams. A weakness in the immune system, poor blood flow in the intestines and too large feedings are possible causes, and common signs of infection are high levels of red blood cells, low oxygen, low heart rate and apnea which is a sign of infection.

The prevalence of these complications are 8.12%, 3.72% and 2.38% respectively in the cases retained for experiments in this thesis. Rare outcomes of interest are typical of medical databases; therefore an abundance of data is needed to obtain a significant number of cases having the rare outcome. This is why many studies use databases that have information from a group of collaborating hospitals, where the data has been collected over a period of years.

2.2 Databases

Medical databases can become very large due to the vast amount of data created by bedside monitors, test results and patient charts. This data requires extensive pre-processing before it can be used in a CDS system because it contains a majority of incomplete cases, outliers, typographical errors, missing data, codes which may be hospital specific and values which are difficult to interpret and may be misleading. For example moribund babies may have a normal Score for Neonatal Acute Physiology (**SNAP**), but they are not healthy babies: they are given comfort care but are not actively treated. Such cases may have values in some fields recorded as "0" (which signifies a normal range) because it is not measured. Similarly, not all physiological parameters may have been recorded in very sick babies due to a lack of time or resources. It is therefore inappropriate to assume that all missing values are normal [Richardson et al. 2001].

2.2.1 Canadian Neonatal Network Database

The Canadian Neonatal Network™ **(CNN)** is a group of multi-disciplinary Canadian researchers from 27 hospitals and 16 universities who collaborate on research issues relating to neonatal and perinatal care. Data collected by the CNN between January 8th 1996 and October 31st 1997, contains admission data from 17 NICUs, which is 75% of all tertiary-level NICU beds in Canada [Richardson et al. 2001]. This database contains 20488 admissions during the collection period, for which data was recorded on day 1 (admission to the NICU) and days 3, 14 and 28 (or discharge).

2.2.2 EPIC and CHEO Databases

The Evidence-based Practice Identification and Change (EPIC) database is a small repository of cases which have been submitted by a tertiary care center implementing the EPIC Process. These guidelines are derived from evidence-based medicine to reduce the incidence of severe lung disorders and hospital acquired infections, and may differ from those applied to cases from the CNN [Lee 2006]. MIRG acquired the 2002 EPIC database which contains 59 cases. It will be used as a further validation set for best models to observe their generalizability and allow a comparison with previous MIRG models.

To further test the generalizability of the final models in the clinical setting, 60 charts which represented atypical or difficult cases which vary in diagnosis, focusing on cases involving respiratory distress were manually abstracted by a master's student with formal training in medical records management, from the archives at CHEO. The student met with neonatologists to receive direction on data collection, because some variables are available from multiple sources with varying accuracy. Cases were not rejected due to missing variables to accurately represent 'real cases' which are primarily incomplete. Missing values in these smaller validation databases will be replaced using the same methodology as for the CNN cases.

2.3 Clinical Decision Support

A CDS system is "a computer based tool using explicit knowledge to generate patient specific advice or interpretation" [Ramnarayan & Britto 2002]. Decision support tools have been developed to address issues such as the limitation of human memory, the growing number of published journal articles, the increasing amount of medical data, the limitations in knowledge of a specialist about other regions of interest, the limited number of specialists themselves, and the great quantity of data that can be collected at the bedside.

In the healthcare field, clinical information systems are expanding from administrative uses (billing, appointment reminders, records or data storage) to more "subtle" uses that may lead to improved clinical outcomes like accuracy of diagnosis and reduced adverse drug interactions. Some types of CDS systems are designed to ease difficult decisions for clinical staff while others are utilized by patients with the aim of promoting positive healthcare results and patient satisfaction. These systems have already shown promise for reducing errors in many types of healthcare environments: broad applications like the triage of patients and specific ones such as the classification of EEG signals [Courtright et al. 2001], [Fu 1994, p.18].

Decision support tools can address complex decisions as well as decisional conflict (choosing between treatment alternatives). They are becoming popular aids for decisions where the options and results are difficult to quantify and the quantity of information becomes difficult to manage, which is especially common in healthcare decisions. Generally, they can be divided into three categories:

1) **Medical education applications**: to supply medical students with case-based examples, to supplement clinical experience (as it may vary from student to student) and to present scenarios and later performance feedback with the aim of increasing their confidence and knowledge [Pyper 2007].

2) **Decision support for patients**: to provide information, empower patients to be active in their treatment decisions, health education and decisional conflict resolution [O'Connor et al. 2002], [Frize et al. 2005].

3) **Hospital-based decision support**: assist doctors in dosing, diagnosis, preventing adverse drug events, alert management and issuing warnings, present patient trends and test results, and display the electronic health record or patient specific outcome prediction [Ramnarayan & Britto 2002], [Catley & Frize 2003].

While MIRG is developing systems in all three categories, this thesis focuses on the latter type (3). Current CDS systems offer mainly diagnostic support, like the *DXplain* system which uses probabilities and heuristics based on disease profiles, or the fuzzy logic approach that was used to differentiate between the "true" and "false" alarm on neonatal pulse oximeters [Berner 1998]. Progress in techniques like data mining and the availability of large medical databases has led to the emergence of decision support designed to produce patient specific administrative and diagnostic information. CDS systems are well poised to help manage that knowledge for clinicians to prevent information overload and make evidence-based medicine more cost effective.

The NICU has been outlined as "an area where medical knowledge and ability has grown dramatically, and where information and communication technology holds enormous potential" [Safran 2003]. A CDS is a technology that presents itself with strong information sorting and managing abilities that can be integrated into the clinical workflow and hospital informatics systems to provide patient specific advice.

2.4 Outcome Prediction

2.4.1 Group Prediction

Outcome prediction, whether for a group or individuals is an important goal of CDS systems. Scoring models are the principal group prediction method. They are developed for various reasons: to assess vitality, severity of illness, and compare the performance of NICUs which may have differing populations. The most commonly used scoring systems are:

The Clinical RIsk for Babies score (**CRIB**): uses three physiologic variables plus birth weight, gestational age and the presence of congenital anomalies to calculate a score for babies weighing less than 1500g. It was part of the first generation of newborn illness

severity scores but was found to be too difficult to apply to babies born outside of a hospital [Richardson et al. 2001].

The Neonatal Therapeutic Intervention Scoring System (**NTISS**): adapted from an adult model, this system uses a physician's responses to illness severity (supplemental oxygen, blood tests, operations etc.) to measure the intensity and complexity of therapies and interventions used in the first 24 hours, weighted according to invasiveness and cost [Richardson et al 1993].

The Apgar: score was first proposed in 1953 by Dr. Virginia Apgar to assess the health of newborns. It was designed to be easy to observe, not require any special equipment and be performed by delivery personnel without difficulty. The APGAR acronym (activity, pulse, grimace, appearance and respiration) was established in 1962 by two paediatricians. The score provides a "valuable criteria for newborn resuscitation" and since neonatal mortality was first found to be strongly correlated with a low (<7) Apgar score at 5 minutes in 1963, most NICUs began recording it [Finster & Wood 2005].

The Score for Neonatal Acute Physiology (**SNAP**): was developed specifically for neonatal care. SNAP was developed to address restrictions of previous models and is a measure of the amount of deviation from normal of physiologic parameters observed at their 'worst' during the first 24 hours of life. It was adapted from two scoring systems which were validated in the adult intensive care. The 37 input variables can be measured in 5-15 minutes and were not found to be correlated to birth weight. SNAP is correlated with physicians' estimated mortality risk, length of stay and nursing workload. The high correlation between SNAP and NITSS reflects that sicker babies tend to require more therapeutic interventions and that SNAP is predictive of total hospital costs.

Richardson et al. [1993] validated the SNAP-Perinatal Extension (**SNAP-PE**) which includes the original 37 SNAP inputs in addition to the baby's birth weight, small for gestational age (**SGA**) status (**SGA** $<5^{th}$ percentile) and Apgar score as a scoring system to observe groups of patients.

Richardson et al. [2001] used logistic regression to reduce SNAP to six key variables linked to mortality in addition to the known factors of birth weight, Apgar score and the small for gestational age status. The variables are: lowest blood pressure, temperature, blood serum pH and PO_2FIO_2 ratio, presence of multiple seizures and urine output. These variables, along with birth weight, SGA and Apgar score at 5 minutes are known as the nine **SNAPPE-II** variables subset.

Scoring systems have been shown to create usable mortality models for groups of patients, but are not specific enough for individual patient predictions or to make decisions regarding the discontinuation of life support. A branch of the engineering and computing fields called medical informatics is focusing on integrating scoring systems into other modeling approaches which have shown greater predictive abilities [Zernikow et al 1998]. Useful models would provide insight into the relationship between measured parameters and clinical events in addition to these scores.

2.4.2 Individual Prediction Models

Medical knowledge is growing rapidly due to technological advances and research. This may improve clinicians' diagnostic accuracy, while increasing cognitive demands for the assimilation and weighting of increasing quantities of information. CDS systems can process a large amount of data to discover pertinent trends and address the information overload faced by diagnosticians. Medical decisions are also fraught with bias and based on personal experience, which may nor may not be representative of general scenarios though "clinician judgement can be improved through feedback" [Berner 1998], to help weigh the different factors that go into a decision to improve clinical accuracy. Doctors use medical, therapeutic and physiologic factors in their decisional model, and tend to be more accurate than physiologic scoring systems alone. The judgement of doctors has a significant impact on triage (number of beds is limited), on initiation of life support and on the ordering of tests.

Group predictions can also be used to compare the performance between hospitals and determine best practices. Zernikow et al. [1999] state that health care providers as well as hospital administrators could make use of "early and accurate length of stay prognoses of preterm neonates both for economic and organisational reasons". For the

parents of a preterm baby, knowing the estimated length of stay could help them make decisions relating to taking time off from work, alternate accommodations if the hospital is far from home, arrangements for the care of other children during hospital visits, etc. Accurate predictions from a CDS system could help parents, hospital administrators as well as doctors make best use of their time and resources.

2.4.3 Challenges

Medical databases suffer from a number of deficiencies such as missing values and rare outcome problems already mentioned which add to the difficulty in creating high sensitivity models. Morbidity is also important as a predicted outcome because it has life-long economic and emotional implications for the baby and the parents [Stevens et al. 1994]. BPD, for example, is an important outcome because it results in chronic lung disease.

Many types of decisions are made in the NICU. Certain ones, like choosing the settings on equipment are made exclusively by the medical staff. Others like choosing to visit or talking to the baby are formed exclusively by the parents. In this complicated environment, many of the critical decisions are made jointly by a team involving doctors, nurses, the parents and often a social worker. For example, the "withdrawal of life support remains a complex interplay of uncertainty of survival, likelihood of severe morbidity, parental values and institutional policy" [Stevens et al. 1994]. Care providers must ensure that patients are informed about each treatment option's potential harms and benefits. It is agreed that these decisions should be made in accordance with the personal values of a patient, or in this case the parents, who generally wish to be involved [O'Connor et al. 2002]. Decisions made in collaboration with parents will be influenced by the parents' values, therefore two very similar cases could have different outcomes depending on the value placed on life and quality of life.

Complications are much more prevalent in low gestational age babies. In the CNN, for example over 80% of NEC cases were diagnosed in babies of 32 weeks gestation or less. NEC can be prevented for babies at risk with large doses of steroids or treated in early stages if detected; otherwise it requires surgery which can increase the hospitalization in terms of cost and duration.

Models should be generalizable due to the differing population bases, incidence of pre term birth, mortality and morbidity rates among the participating centres. Lee, the director of the Canadian Neonatal Network "reported large variations in practice among Canadian neonatal intensive care units" and that "suitable models for implementing NICU practice change based on evidence and data are lacking" [Lee 2006].

Currently, only 5% of the data in the CNN database is verified for accuracy [CNN 2004]. Subsequent versions of the database at CHEO will contain cases with automatically collected data from electronic patient records and bedside monitors, eliminating the need for time consuming manual entry of cases, as well as some errors that occur during the current data entry process. As such, the next database used by MIRG may contain fewer missing/erroneous values because variables such as blood pressure (13-16% missing in CNN) and percentage of inspired oxygen (63% missing in CNN) would be automatically collected from the blood pressure monitor and ventilator respectively in many centres. In the CNN database, parameters which come from ventilators have some of the highest missing values (oxygenation index 73%, po2fior2 64%, po2 63% and pco2 40%).

A final difficulty related to the development of outcome models obtained from medical data is that there are no known equivalent scoring systems for length of stay, duration of ventilation or important complications like BPD, IVH or NEC to provide a benchmark for comparison or to suggest an initial variable subset. The SNAP variables will provide the starting point because they are measured most frequently and are governed by a strict protocol, as opposed to the NTISS score which could differ between hospitals and over time as protocols evolve [Zernikow et al. 1999].

CHAPTER 3 LITERATURE REVIEW and RELEVANT WORK

3.1 Artificial Neural Networks

Artificial neural networks (**ANNs**) are the most widely used non knowledge-based type of network for non-linear medical outcome prediction applications. They are chosen for their ability to efficiently process large amounts of data which may be very noisy and model complex non-linear interactions between a set of input variables and one or many outputs. Their ability to develop generalizable models and integrate different types of data is also an asset. An ANN is a distributed information processing concept inspired by biological neurons. It can be trained for pattern recognition and data classification because of its proficiency for adaptive learning and self-organization without the need to clarify the link between variables and the output.

The links between the neurons are adjusted as the system develops (learning process) and this feature is particularly useful because a network can be re-trained as the input patterns evolve: new disease trends, tests and known risk factors. Modeling physiological processes is often more successful with non-linear approaches because the processes tend to be non-linear [Baxt 1994]. ANNs provide some advantages over certain statistical and rule-based systems because no prior knowledge of the relationship is needed, the inputs require less pre-processing and in many cases they have detected previously unknown key factors [Ennett et al. 1999], [Frize, Ennett & Charette 2001].

ANNs can be separated into four basic computational models [Fu 1994], [Baxt 1994], [Bishop 1995]:

Classification: Primarily single and multilayer perceptrons, classification models assign the input data to one of a finite number of categories. Sections 3.1.1 and 3.1.2 detail two types of classification-based ANNs in greater detail.

Association: Hopfield networks are a common type of associative network. Used for autoassociation and optimization tasks, they are based on the concept of energy surface minimization applied to binary-valued neurons [Fu 1994]. These networks are often integrated into VLSI chips and can serve as content addressable memory systems.

Optimization: Boltzmann Machines are referred to as the stochastic (as opposed to deterministic) version of Hopfield networks. They also use the concept of energy minimization to move towards the global minima but an element of randomization is introduced to help the network escape local minima.

Self-Organization: Models such as Kohonen Networks and Hebbian learning have the ability to lean and organize information without being provided with the correct output for the given input data. This is known as 'unsupervised learning', and differs from the supervised training method used in this thesis' classification models where each training case is associated to a desired output. These networks are often used in robotics where the ability for "self-organization compensates for inaccuracies and noise in sensor readings and offers a method for dealing with unexpected and changing situations which lack mathematical descriptions" [Fu 1994].

3.1.1 Mean Squared Error Classification ANNs

The performance of a mean squared error (**MSE**) classification neural network application hinges on how well it can classify the outcome for new and unseen cases. Medical databases have the difficulty of not always containing large amounts of consistent data which often results in neural networks that misclassify cases "in low-prevalence classes in order to correctly classify patterns in high-prevalence classes" [Penny & Frost 1994]. MSE ANNs operate by using the steepest descent optimization to reduce the error between the actual and desired output. The following section details important elements about MSE ANNs.

3.1.1.1 Structure

The smallest processing unit in a neural network is the neuron (or node). Nodes are placed in parallel, and the output of a node is the weighted sum of its inputs, affected by the activation function as displayed in Figure 3.1. When a node receives non-null stimuli, it propagates a signal to the following layer of nodes.

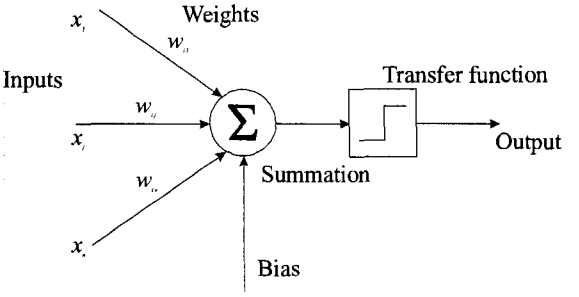

Figure 3.1. Artificial neuron

The hyperbolic tangent function, $\frac{e^{2x}-1}{e^{2x}+1}$ (fig 3.2), is a good choice for an activation function for these experiments because it transitions faster between output values than the sigmoid function, which can lead to faster learning [Bishop 1995, p. 127]. This non-linear function quickly departs from zero in both directions, and when the inputs are scaled to have zero mean and unit variance the learning speed is further hastened [Penny & Frost 1996]. It also serves to limit the activation of the neuron to a continuous range between 1 and -1.

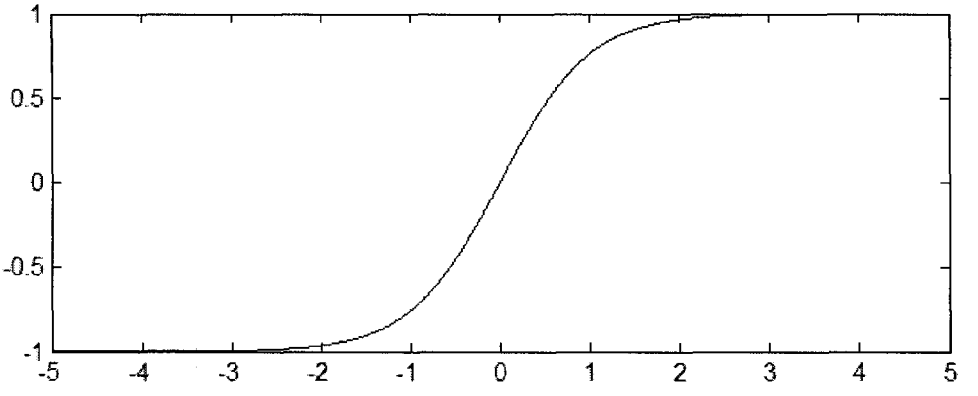

Figure 3.2. The hyperbolic tangent activation function

The input layer encodes the data presented to the network and the hidden layer encodes the non-linear relationship between the inputs and the output(s) [Fu 1994, p.18]. Hidden layers "can form more complex decision regions" and are required to map non-linear relationships [Fu 1994, p.31]. One hidden layer usually gives the network sufficient complexity to approximate any smooth nonlinear function [Penny & Frost

1996]. Figure 3.3 demonstrates the typical network structure of a single output 3-layer network.

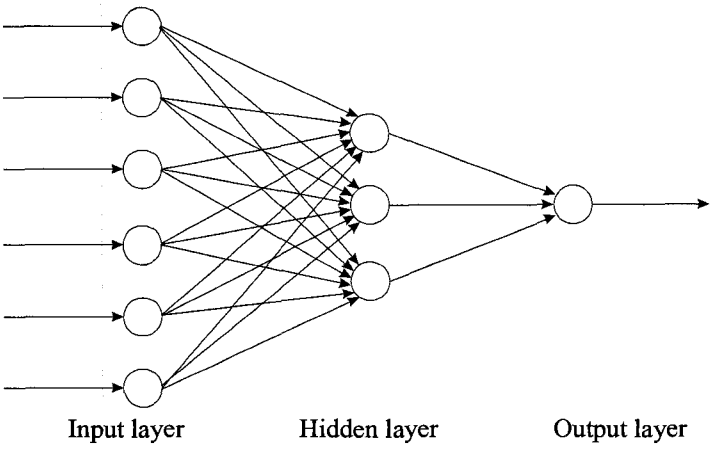

Figure 3.3. Structure of a 3-layer neural network

In the case of medical applications, inputs to the network come from sources such as patient charts, test results and physician observations. The values may be nominal (numbers corresponding to different categories i.e.: diagnosis), ordinal (representing the degree of severity of a symptom i.e.: Apgar score), interval (i.e.: gestational age range) and continuous (i.e.: heart rate). In the output layer, n nodes are used to represent n outcomes. Structures in this thesis contain one output node, and its activation represents the presence of the underrepresented outcome under investigation.

Fitting to noise, or over fitting, occurs when the complexity of the network exceeds that of the function to be estimated, and the network learns or memorizes features about the training data that are not present in the test and validation sets [Penny & Frost 1996]. Memorizing is often manifested by good classification results of training data and poor classification results of validation sets. Weight elimination and variable reduction are introduced later as a way to reduce noise, improve the classification rate and simplify the network.

Structures

Structures used to model non-linear data have at least one hidden layer with two or more hidden nodes. Factors affecting the number of hidden nodes include the number of inputs and outputs, the number of training cases, as well as the complexity of the data

itself. The development of a model with satisfactory performance may be time-consuming; however once in a clinical setting the network will not need to be retrained frequently and can be done while the current version is still operational.

A few mathematicians and researchers have investigated the relationship between hidden structure and performance. Certain researchers, such as Livingstone and Manallack, Kolmogorov, and Masters have tried to set bounds for the largest or smallest structure capable of adequately modeling a non-linear function.

Livingstone and Manallack's [1993]: Their equation determines the maximal number of hidden nodes that will result in a network with acceptable generalization using the ratio of cases to connection weights. Their equation for determining the maximum number of weights (w) is the most prominent method using the number of training cases (tr) and the number of outputs (o) as variables. For good generalisation, the number of weights is chosen to make d $\geq$3.

$$d = \frac{tr * o}{w} \tag{3.1}$$

The training datasets used in this thesis initially consist of 27 input variables and over 7000 cases, which corresponds to an upper limit of 84 hidden nodes. This upper limit quickly becomes unmanageable as variables are removed because the datasets used are so large. It is the only method found that recommended an increasing number of hidden nodes when variables are removed. The maximum number of hidden nodes determined by this method is the largest of all methods/theorems considered.

Kolmogorov's Theorem (Universal Approximation Theorem): Kolmogorov showed that every function of n variables can be represented by a superposition of a small number of functions of one variable. Extended to neural networks, this theorem says that n input variables can be "represented exactly" with 2n+1 hidden nodes (n is the number of input variables), meaning that structures with 2 to 2n+1 nodes in the hidden layer should be attempted in order to obtain the best possible model. The number of structures to attempt would decrease at the number of input variables is reduced. The maximum number of hidden nodes decreases at twice the rate of the variables [Bishop 1995, p. 138].

Masters Theorem: According to Masters, the ideal number of hidden nodes decreases when variables are eliminated from models, but at a much slower rate than with Kolmogorov's theorem. Masters' [1993, p. 177] defines the number of hidden nodes by equation 3.2 where h is the number of hidden nodes, i is the number of input nodes and o is the number of output nodes: $h = \sqrt{i * o}$ (3.2)

While Kolmogorov's theorem finds an acceptable structure by varying the number of hidden nodes, Masters suggests keeping a stable structure and instead varying the network's parameters to obtain a model which meets set performance criteria.

It is impractical to use Livingstone and Manallack's upper bound because such structures could take weeks to converge and it increases the risk of overtraining. This offers little additional classification advantages during the variable elimination process where the network structures will have to be re-run multiple times. Practically, the network's performance can be more precisely tailored by adjusting parameters in the stopping criteria and using weight tuning methods than by adding extra hidden nodes or layers after a certain point. Masters theorem may be applicable to datasets consisting of uniform cases with a low level of complexity. The outcomes modeled in this thesis may be too complex to be adequately modeled by the small number of nodes suggested by Masters. Kolmogorov's approach for finding an appropriate model seems to be the best choice for this work.

3.1.1.2 Control Parameters

The MSE ANN developed by MIRG contains nine control parameters that modify the connection weights and biases of the network and can be tailored to improve the performance of the model.

Learning Rate (lr, lr_inc, lr_dec): The value of the learning rate determines the speed at which the network attains a minimum in the criterion function (average squared error). It must be chosen wisely to avoid oscillation around a local minima and lack of convergence. During training, the learning rate will be incremented or decremented depending on the error ratio.

Momentum (m): The momentum is a key variable in optimizing network performance. It is a proportion of the previous weight change that is added to the new value to give the gradient descent algorithm inertia so that the error value does not get caught in local minima. It also helps to counteract slow learning which is often associated with the back-propagation algorithm [Fu 1994, p. 57]. It allows the network to ignore small features in the error surface [Frize et al. 2000]. A smaller momentum complements a larger learning rate and vice versa. Equation 3.3 shows the weight update equation when the momentum term is used. The fraction term m, ranges from 0 to 1. It will change according to a comparison of current-to-previous iteration errors. When the momentum is set to zero, current weight change operates without the momentum. Otherwise, current weights and biases are calculated according to the modified delta rule. When the momentum is one current weight change, $\Delta W_{ji}(n)$, is equal to previous weight change and the gradient, $\delta_j(n)$, resulted from that is ignored [Bishop 1995, p. 267].

$$\Delta W_{ji}(n) = m\Delta W_{ji}(n-1) + lr * \delta_j(n) y_i(n) \tag{3.3}$$

Weight Decay Constant (lambda, lambda_inc, lambda_dec): The weight decay constant determines how strongly the weights are penalized. It limits the size of the connection weights by penalizing large weights by stopping them from increasing, thereby reducing the variance of the weights in a network [Frize et al. 2000]. It is also incremented or decremented depending on network performance.

Weight Elimination Scale Factor (w_0): The weight elimination scale factor defines the size of large and small weights. Weights that are smaller than the weight scale factor are forced to zero so that they are removed from the network. Weights larger than the weight scale factor are not eliminated but may be penalized using the weight decay constant [Frize et al. 2000]. The weight elimination scale factor is significant for variable reductions. The optimal value of the scale factor is difficult to determine without seeing the results of multiple network iterations because the range of the weights and biases may vary between structures and models. Both the weight decay constant and the weight elimination scale factor are incorporated into the weight-elimination cost function described in section 3.1.1.6.

Error Ratio: The error ratio value controls how the back propagation makes adaptive changes in the learning rate, weight-decay constant and the momentum. The error ratio is used to test the error from one epoch to the next, with the error ratio value being the improvement required for the training to continue. If the training set error from the current epoch is larger than the previous epoch's multiplied by the error ratio, the learning rate is decremented, otherwise it incremented.

Charette and Rybchynski greatly improved MIRG's MSE ANN performance by automating optimal parameter selection [Frize, Ennett & Charette 2001], [Ennett et al. 2004]. The search for the parameter values resulting in the highest performance for each network parameter occurs over pre-defined ranges. The algorithm searches for the two optimum performance points, "calculates their midpoint, and then selects two new maximum points from these points. This is repeated recursively until the network converges upon the network parameter value that gives the maximum network performance" [Rybchynski 2005].

3.1.1.3 Data Resampling

Neural networks trained on highly skewed outcomes tend to produce models which classify the rare outcome quite poorly and require a long time to train [Frize et al. 1998]. This fact prompted Ennett & Frize [2003] to investigate methods capable of rectifying this imbalance. Changing the *a priori* distribution of the training set can be accomplished by removing cases from the prevalent outcome or increasing cases from the rare outcome. The latter approach was taken to avoid the loss of information that would arise from removing cases. The prevalence of the rare outcome was increased to 20% in the training set of each outcome by artificially resampling cases, which helped reduce training time. "The training set is used for learning and adaptation, performed by the adjustment of weights during training. The results from the testing set determine when it is possible to stop training" [Catley 2007, p.22]. Evaluating a model with a validation set with the actual *a priori* distribution and cases which were not used to train or test the network provides the best estimate of the generalization error [Dreiseitl & Ohno-Machado 2002], [Ennett 2003].

3.1.1.4 Normalization of Data

In an effort to reduce training time, inputs are normalized using the z-score formula of equation 3.4, which is also called linear rescaling. The z_n is the normalized value of variable x, x_n is the original value of the variable and X and σ are the mean and standard deviation of the variable x [Olden & Jackson 2002]. The outputs are set to -1 (common outcome) or 1 (rare outcome) [Bishop 1995, p. 278].

$$z_n = \frac{x_n - X}{3\sigma} \tag{3.4}$$

3.1.1.5 Learning and Weight Updating

The multilayer perceptron learns in two steps: first there is the forward propagation of inputs to produce the networks output, and then the backwards propagation of the error to adjust the weights. The forward pass of the inputs to the output of the hidden nodes is computed as follows: the inputs (x_i) are weighted (w_{ji}) and then summed together with an added bias term b_j. This value is then transformed by the activation function of the hidden node, g(), and passed to the output layer as in equation 3.5, where z_j is the output of the j^{th} hidden node receiving d inputs [Bishop 1995, p.118].

$$a_j = \sum_{i=1}^{d} w_{ji} x_i + b_j \quad \text{and} \quad z_j = g(a_j) \tag{3.5}$$

An analogous formula applies for the forward pass of the hidden node outputs to the network output. Once the forward computation is complete, the output error is calculated. Back propagation is carried out using the delta rule represented by equation 3.5 to change each weight for the next iteration. "The basic idea behind the back-propagation algorithm is to use gradient-descent move in error space with every iteration, so as to decrease the error faster and search for the point in the error space with a (global) minimum of the mean squared error" [Zhou 2006]. The back-propagation algorithm computes as follows for evaluating the derivatives of the error E^n, which is the sum of the error over all training patterns, with respect to the weights in four steps:

Step 1: apply an input vector and propagate forward to the output node to determine the activation of the hidden and output neurons using (3.5).

Step 2: evaluate the δ_k, the local descent, for the output unit.

$$\delta_k = g'(a_k)\frac{\partial E^n}{\partial y_k} \tag{3.6}$$

Step 3: back propagate the derivatives, δ, to obtain a δ_j for each hidden unit

$$\delta_j = g'(a_j)\sum w_{kj}\delta_k . \tag{3.7}$$

Step 4: calculate the derivatives using:

$$\frac{\partial E^n}{\partial w_{ji}} = \delta_j z_i \tag{3.8}$$

For the batch training algorithm considered in the MSE ANN configuration, the delta rule dictates the changes in the weights as:

$$\Delta w_{ji} = -\eta\sum \delta_j^n x_i^n \tag{3.9}$$

where η is the learning rate. The learning rate and momentum parameter described later are used to determine how quickly each step approaches the minimum. The gradient is calculated with respect to each weight.

3.1.1.6 Weight-Elimination and Variable Elimination

Many methods have been developed to avoid overfitting. They fall into two categories: restricting the model complexity, or restricting the influence of the data on the network parameters. Network pruning falls into the latter category. It is the process of removing some weights, which can be done without negatively affecting the network performance in a fully connected network [Fu 1994, p.92]. Simplifying networks by removing variables has the advantage of making more accurate networks in terms of sensitivity. Variable elimination can be performed one input at a time to identify which inputs act as noise and hinder, rather then help, classification. It also leads to better generalization and quicker training time because variables which are not indicative of the desired outcome are removed and each remaining variable will have a more pronounced influence on the training [Fu 1994, p.92-93].

While weight decay penalizes larger weights and creates outputs with less variance, weight elimination reduces the small weights to zero and thus their effect on the network to zero [Trigg 1997]. "Weight decay and weight elimination work best when using a large initial network structure, small initial weights and a small learning rate" [Frize et al. 2000]. The weight elimination function, as provided by Weigend et al. [1990] is shown in equation 3.10. It aims to solve the overfitting problem by adding a complexity term to the cost function.

$$E(W) = E_0(W) + \lambda \times \sum_{ij} \frac{\frac{w_{ij}^2}{w_0^2}}{1 + \frac{w_{ij}^2}{w_0^2}} \qquad (3.10)$$

Where:

E_0 = SSE (sum squared error).

λ = weight decay constant as described in 3.1.1.2.

w_{ij} = weight of connecting node i and j.

w_0 = weight scale factor, as described in 3.1.1.2.

The weight decay terms limits the magnitude of the weight to make the decision boundaries smoother and makes it harder for the network to memorize particularities in the training set [Dreseitl & Ohno-Machado 2002].

3.1.1.7 Performance Measures

Correct Classification Rate: The correct classification rate (**CCR**) represents the total number of cases that were classified appropriately by the MSE ANN. The constant predictor (**CP**) is the highest number of cases predicted correctly if a constant value is used as a predictor. For example, if the event of mortality had a 5% prevalence in a database, a constant predictor would correctly predict 95% of the cases.

Sensitivity and Specificity: The confusion matrix is a means to analyze the classification performance of a system. It displays the classified outcomes into four categories

depending on their predicted and actual outcome. From the confusion matrix, values of sensitivity and specificity can be derived. In this work, a positive outcome will be the rare outcome (i.e. mortality) therefore a true positive is a non-survivor correctly classified as a non-survivor and a false negative is a non-survivor incorrectly classified as a survivor. Sensitivity is the rate at which non-survivors are correctly classified. Specificity is the rate at which survivors are correctly classified. The ideal value for sensitivity and specificity is 1. Sensitivity and specificity are affected by the prevalence of each outcome. Sensitivity primarily obtains its value from the true positives (20% of cases in training sets and 2-8% of cases for highly skewed outcomes), while the value of specificity relies on the true negatives, or 80% of the cases in the training set and 92-98% of cases. The MSE ANN will naturally tend towards higher specificity, which is why alternate stopping criterions were investigated.

		Actual Outcome	
		Negative (survivor)	Positive (non-survivor)
Predicted Outcome	Negative	True Negative (TN)	False Negative (FN)
	Positive	False Positive (FP)	True Positive (TP)

Figure 3.4. A confusion matrix for two-value classification outcome

$$sensitivity = \frac{TP}{TP + FN} \qquad (3.11)$$

$$specificity = \frac{TN}{TN + FP} \qquad (3.12)$$

When dealing with rare outcomes, the network performance, if measured by the classification rate, can be misleading. This is why sensitivity and specificity are used instead.

The cost of misclassification is not the same for false positives as for false negatives. Misclassifying can have a potentially significant impact when dealing with medical databases because the rare outcome often mandates action in the course of treatment. Using sensitivity and specificity allows the network to be evaluated in terms of how accurately each outcome is classified. Aiming to maximize the specificity reduces the number of false positives, a survivor misclassified as a non-survivor.

Receiver Operating Characteristic (ROC) Curve: The receiver operating characteristic (**ROC**) curve is a graph of sensitivity versus 1- specificity, also described as true positive fraction versus the false positive fraction. The area under the curve gives a "definitive measure of the classifier's discrimination ability that is not dependent on the choice of decision threshold" [Penny & Frost 1996].

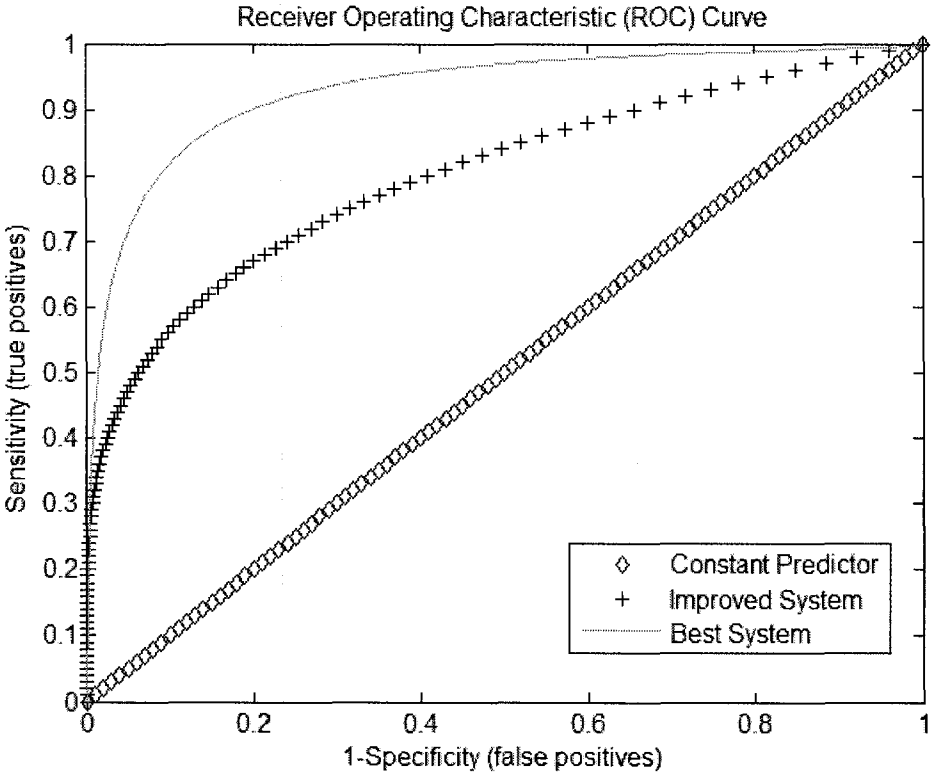

Figure 3.5. Receiver operating characteristic curve

The ROC is a graph of the true positive rate versus the false positive rate to demonstrate the discriminating ability of the network over a range of conditions. ROC curves can be used to compare predictive systems. When looking at one specific ROC curve, as on Figure 3.5, moving towards the right corresponds to over-diagnosing, while moving to the left reduces the false positives but means that fewer patients will be diagnosed even though they are positive.

The "Constant Predictor" from Figure 3.5 represents random guessing based on the known prior probabilities of each outcome. A system whose performance equals the constant predictor offers little over educated guessing for balanced outcomes. The "Improved System" is superior because it achieves higher sensitivity with a lower rate of

false positives. The most desirable in outcome prediction, "Best System", is the curve that remains closest to the y-axis as the y-value increases because it represents the system that would correctly diagnose the largest amount of true positives while limiting the false positives. For the purposes of this thesis, models with an area under the ROC ≥ 0.85 will be considered acceptable.

Logarithmic Sensitivity Index: The logarithmic-sensitivity stopping criterion is introduced to help balance the rarity of some outcomes and improve the correct classification of the rare outcome. It finds a balance of sensitivity and specificity while slightly favouring a higher sensitivity. The *logsens* tends towards infinity as the sensitivity and specificity tend towards one. The n value can be modified to put additional emphasis on sensitivity ($0 \leq n \leq 1$) [Ennett et al. 2003].

$$logsens = -sensitivity^{n} * \log(1 - sensitivity * specificity) \qquad (3.13)$$

This stopping criterion has been shown to perform equally or better than the minimum mean squared error or the correct classification rate alone in databases that present highly skewed outcomes [Ennett et al. 2003].

Using an early stopping criterion is a common way to increase the model's performance in testing and validation sets as it limits the model's ability to adapt to the training set (limits memorization) [Dreiseitl & Ohno-Machado 2002], [Bishop 1995, p. 343].

3.1.1.8 Relative Weight Calculation

Medical databases contain a multitude of variables collected from the patient, and therefore it is important to reduce this extensive list to key factors. Relative weight calculation is a process where the weights and biases are converted to a value, usually between 0 and 100, to express their importance in estimating the output. It can provide guidance on which variables to remove to obtain this minimal variable set. In linear networks, converting the weights to a value [0,100] can be done in a single step. In non-linear networks, like the ones presented in this thesis, the conversion process is more complicated due to the presence of weights and biases in multiple layers. One approach

is to remove one variable at a time and observe the change in the networks performance, but this approach is very time consuming and does not provide a direct ranking of the variables.

Rybchynski [2005] researched alternatives and explored the Garson-Goh weight extracting method [Goh 1995].

Rybchynski [2005] adapted the original algorithm for use in structures with hidden layers. The following sets of equations demonstrate the relative importance calculation using weights and biases of a network, where:

j: is the number of hidden nodes

k: is the number of output nodes

i: is the number of input nodes

w_{ij} : is the connection weight between nodes *i* and *j*

Step 1: Calculate the relative weighting of the hidden-output connection for the output node: $rw_{jk} = \frac{abs(w_{jk})}{\sum_{j} abs(w_{jk})}$ (3.14)

Step 2: Calculate the relative weighting of the input-hidden connection for each hidden node: $rw_{ij} = \frac{abs(w_{ij})}{\sum_{i} abs(w_{ij})}$ (3.15)

Step 3: Multiply each relative weighting input-hidden connection value by the relative weighting value of its connected hidden-output weight:

$$P_{ij} = rw_{ij} * rw_{jk} \quad (3.16)$$

Step 4: Sum the final relative weighting products: $S_i = \sum_{j} P_{ij}$ (3.17)

Step 5: Divide the sum of the final relative weightings for each input node by the total of all relative weightings: $RI_i = \frac{S_i}{\sum_{i} S_i}$ (3.18)

Rybchynski wrote a MatLab program to convert the weights and biases from the fully-connected non-linear networks to a single value per variable. This program was

modified to take these values and expand them to the [0,100] interval so that they can be easily ranked.

Relative weight ranking is also performed to evaluate how well a model correlates with medical experience and to determine the most concise number of variables that provides the best model. An advantage to this is that once a minimal dataset is created, fewer variables need to be imputed in incomplete cases. Initially, when variables are chosen because of medical relevance and database availability, as many variables as possible are included to develop a rich (complex) model [Jerebko et al. 2003]. Variables are then ranked and eliminated in "an attempt to create a network that would be generalizable to other databases" [Catley 2007, p. 108]. Variable elimination helps deal with problems of pattern recognition in a high dimensional space [Bishop 1995, p. 130].

3.1.2 Maximum Likelihood Classification ANNs

An alternate ANN configuration to the MSE ANN described earlier was developed by MIRG researchers for the purpose of developing an estimated risk stratification for rare outcomes. The maximum likelihood (**ML**) estimation was applied to approximate a neural network probability function by changing the network parameters, specifically: the number of hidden nodes, the weights, and the biases. "Essentially, this means that the cost functions in neural network training (such as the MSE from the MSE ANN described in section 3.1.1) are modeled as maximum likelihood functions in order to *estimate risk stratification*" [Zhou 2006], [Catley 2007]. "Traditional neural networks based on the gradient descent search on the error surface to seek a point of the minimum of the global error. In our case, on the contrary, the global maximum of the likelihood function is the aim of the neural network learning. Therefore, a new weight update technique with the gradient ascent search for two and three layer neural networks is desired" [Zhou 2006].

3.1.2.2 Weight Update

In his development of a new training algorithm Zhou [2006] defined the training data D of (independent) patient records as $\{<x_1, y_1>, <x_2, y_2>.....<x_m, y_m>\}$, where x_m

represents the values of a set of input variables for the m^{th} patient and y_m corresponds to the outcome, which takes values of 1 or 0 (mortality or survival).

The likelihood function L(θ), where $f(<x_i, y_i>; \theta)$ denotes the probability of y_i, given the input as x_i in the i^{th} patient record is given as:

$$L(\theta)=f(<x_1, y_1>, <x_2, y_2>.....<x_m, y_m>;\theta)= \prod f(x_i, y_i|\theta), \qquad (3.19)$$

Zhou [2006] links the conditional probability that $y_i=1$ given x_i is $P(y_i=1|x_i)$, to the neural network output $O(x_i)$ with $P(y_i=0|x_i)=1-O(x_i)$, and rewrites equation 3.19 as equation 3.20:

$$\prod_{i=1}^{m} O(x_i)^{yi}[1-O(x_i)]^{(1-yi)} \quad (3.20)$$

Which is re-written as equation (3.21) in a log-likelihood form:

$$L(\theta) = \ln[L(\theta)] = \sum_{i=1}^{m} y_i * \ln O(x_i) + (1-y_i) * \ln(1-O(x_i)) \quad (3.21)$$

Equation (3.21) is the likelihood function that is also the cost function of the alternate neural network configuration. The training process finds a set of parameters (hypothesis) that maximize the likelihood function in the parameter space (hypothesis space) [Zhou 2006]. The ML estimation is based on two conditions: the samples in the training set are considered independent each other and the training set is sufficiently large to ensure the ML estimation will converge to the true probability.

Mitchell [1997], wrote the ML hypothesis as :

$$h_{ML} = \arg\max_{h \in H} \sum_{i=1}^{m} di \ln O(xi) + (1-di)\ln(1-O(xi)) \qquad (3.22)$$

"The right hand side of the equation (3.21) is rewritten as a simple form G(O, D). Now, we can use gradient ascent (i.e. the learning rate is positive) to derive neural network weight updating rule so as to maximize the G(O, D)", where *di* is the desired signal [Zhou 2006]. Mitchell [1997] presented the derivation of the weight-updating rule

for a single layer neural network, which is also valid for output neurons in multi-layer networks:

$$\Delta W_{jk} = \eta \sum_{i=1}^{m} (di - O(xi)) x_{ijk} \qquad (3.23)$$

x_{ijk} is the synapse input from hidden neuron k to output neuron j when a neural network is taking the i^{th} training pattern. The weight update equation for hidden neurons is similarly obtained.

Zhou [2006] derived the weight training algorithm to use the ML estimation as a cost function in the MIRG framework for compatibility and continuity with the MSE ANN. The ML ANN uses the sigmoid function, $\frac{1}{1+e^{-x}}$, as the activation function because its output ranges from 0 to 1.

3.1.2.3 Additional Performance Measures

Two important performance evaluation parameters of the ML ANN model are the ROC curve and the Hosmer-Lemeshow (**H-L**) goodness-of-fit test. The H-L test does not measure the discrimination ability of the model like the ROC Curve, but rather how well the networks estimated risk stratification reflects the data. This is also known as the calibration of the model and evaluates the correspondence between the observed and expected number of the outcomes over the entire probability range ([0,1]). The H-L test divides all data into 10 groups and calculates the number of expected and actual rare and common cases classified in each group. The Pearson Chi-square statistics $\hat{C}$ (equation (3.24)) is used across the groups to measure the calibration where n_k'is the number of patterns in the k^{th} group, $\bar{\pi}_k$ is the average estimated probability and o_k is the number of observed rare outcome cases (i.e.: deaths) in the k^{th} group.

$$\hat{C} = \sum_{k=1}^{g} \frac{(o_k - n_k' \bar{\pi}_k)^2}{n_k' \bar{\pi}_k (1 - \bar{\pi}_k)} \qquad (3.24)$$

Zhou suggested that "any group with the expected deaths no greater than 5 would be merged with one of its neighbour groups, since the small number of expected outcomes would bias the P-value and affect the assessment of the model's goodness of fit" [Zhou 2005]. Models with a P-value > 0.05 are deemed valid. A higher P-value indicates a greater goodness of fit of the estimated risk stratification compared to the actual statistics [Richardson et al. 2001]. The degree of freedom represents the number of risk categories minus two, therefore a value of 8 represents 10 risk categories of a width of 0.1 out of the [0,1] interval.

Sensitivity and specificity still apply with the difference that the positive outcome of death (or presence of a complication) is still represented by 1 and survival by 0 instead of -1. Two approaches for improving sensitivity without modifying the prior probabilities were studied; adding a scalar term to the cost function, and adjusting the classification cut-off point of an existing model.

The first approach uses the weight $0< \alpha <=1$, to lessen the effect of the error vector of the dominant outcome category. The new cost function is represented as E, where E(w1) is the error caused by the rare outcome and E(w2) is the error caused by the dominant outcome [Zhou 2006]:

$$E(w)=E(w1)+ \alpha E(w2) \qquad (3.25)$$

Guiding the error towards the rare class results in a network that learns more from the rare class, thus increasing the sensitivity. Zhou [2006] tried this approach and, though the sensitivity increased, failed to obtain an acceptable probability estimation model based on the result of the H-L C-statistics. The method was therefore unacceptable.

The second method, adjusting the cut-off value, is applied once a model has been created. Zhou created a mortality model with the 13 key indicators identified by Ennett [2003], and gradually decreased the classification cut-off point from 0.5 down to 0.15. The results improved the sensitivity even when corrected for the drop in specificity. His results are shown in table 3-1.

Table 3-1. Effect of varying cut-off point on sensitivity of a model [Zhou 2006]

Cut-off	0.5	0.4	0.35	0.25	0.15
Sensitivity %	25.2	33.5	37.8	47.8	59.1
Specificity %	99.3	98.7	98.3	97.7	95.9
CCR %	96.7	96.4	96.2	95.9	94.6

3.2 Case-Based Reasoning

Case-based reasoning is an analogy approach that makes use of the mathematical similarity between cases to solve problems. In medical applications, CBR can be used to compare past patients to a current patient who has similar symptoms and use a repository of past cases to diagnose the current patient. The basic concept is very different from neural networks because the approach seems intuitive and transparent. The primary assumption is that the matching database contains cases of sufficient complexity and richness [Aamodt & Plaza 1994].

MIRG has used case-based reasoning to match closest cases for inspection by a clinician, predict outcomes, extend datasets to include missing minimum data set variables and impute missing variable values into patient cases. As stated by Aamodt and Plaza [1994], the CBR cycle consists of the following four steps:

1) Retrieve the most similar case(s) from the complete case database

2) Reuse the information in the case(s) from 1) to solve the current problem

3) Revise the proposed solution

4) Retain certain features of the problem/solution

This thesis limits the use of CBR to the first two steps of the process to impute missing values.

3.2.1 Replacing Missing Values

Missing values are commonly encountered in databases used in scientific and engineering analysis. In some cases the missing values and the cases which contain them can be removed from a dataset, however in many cases this would result in a loss of information and these missing values are instead imputed. This thesis deals with missing value replacement and parameter estimation techniques which are applicable to, and in

fact used in many fields such as the estimation of missing parameter estimation in DNA microarrays [Troyanskaya et al. 2001], location estimation using the Global System for Mobile communication with multiple signals in rough terrains [Wu et al. 2007] and the replacement of missing values in surveys used for market analysis [Noquiera et al. 2007] to name a few. The specific case of the CNN medical database is used in this work to illustrate the use of imputation methods and ANNs.

The case retrieval step in the CBR process is often conducted using the k-NN algorithm, which ranks the cases according to similarity and uses the parameter k to determine how many to retrieve. The k-NN distance algorithms plot all variables of a case in k-dimensional space. The distance from the target case to all other cases in the case-base is calculated. The cases with the smallest distance to the target case are chosen as the most similar.

In a weighted k-NN algorithm the axes of k-space are not uniformly set: input variables with higher (lower) weights are plotted on a new elongated (shrunken) axis, forcing the difference in values along the axis to be more (less) important. A strength of the k-NN classifier is that it functions with missing values and therefore it can be used to fill missing values. The variable with a missing value is not drawn on the axis for that particular case and a distance is still calculated. The algorithm works as follows:

i: the number of input variables

n: number of complete cases used in the matching

w_i: the weight of the input variable i

t_i: value of the i^{th} input variable of the incomplete case

x_i: value of the i^{th} input variable of the complete case

Step 1: Calculate the distance between each complete case and the incomplete case.

$$dist(t_i, x_i) = \sqrt{\sum_{i=1}^{n}[(t_i * w_i) - (x_i * w_i)]^2} \qquad (3.26)$$

Step 2: Calculate the similarity and retain the k most similar cases.

$$similarity(t_i, x_i) = 1 - \frac{dist(t_i, x_i)}{\max\_dist} \qquad (3.27)$$

Step 3: Missing values in the incomplete case are replaced with the mean value of that same variable in the k closest cases to create a complete case.

If the relative importance of the inputs is known, the weighing factor w_i can be used to bias the importance of each variable accordingly. In cases where the importance is not known or not required as during an outcome-independent imputation, all w_i on variables are set to 100 and those weighing outputs or administrative fields are set to zero.

Some variables have little variation (ex. Lserum), while others vary greatly (ex: birth weight). For the k-NN to be unaffected by the different range of each variable, the variables were normalized to have the same 0 to 1 range to ensure they have equal effect on the similarity calculation of the k-NN.

In this work, the **Mean (k-NN)** terminology will be used to describe the CBR process where equal weights are used to find the 10 closet matching cases, and a missing value is replaced by the mean value of those 10 cases.

3.3 Relevant Outcome Prediction Models

3.3.1 Performance of Mortality Models and Scoring Systems

An effort at true outcome prediction was made by Zernikow et al. [1998], who used an artificial neural network to classify cases with respect to outcome (mortality or discharge). A neural network was trained with a database of 890 cases of newborns having a birth weight of less than 1500g, with 77 (8.7%) mortalities among the cases. The model with the best classification results had 13 inputs and 3 hidden nodes. The results are summarized in table 3-2.

Table 3-2. Summary of classification results [Zernikow et al. 1998]

	Sensitivity (95% confidence interval) (%)		
Specificity (%)	ANN	Logistic regression model	P Value
75	98.8 (93.1 to 99.9)	94.7 (85.1 to 98.6)	0.061
80	97.5 (89.9 to 99.6)	90.6 (78.8 to 96.7)	0.029
85	94.6 (84.2 to 98.7)	83.5 (69.5 to 92.5)	0.010
90	87.8 (74.2 to 95.4)	70.6 (54.4 to 83.5)	0.002
95	69.6 (51.6 to 83.7)	46.1 (29.9 to 64.0)	0.001

The network was found to underestimate the mortality risk in heavier and more mature babies, but no causes or solutions to this problem were presented [Zernikow et al. 1998]. Physiological measurements like cord blood pH, oxygen requirements in the first 24 hours and the presence of signs of infection were values that were not available but thought to be potentially helpful to classification in future versions of the MSE ANN. Using the same data, they derived a second model using stepwise logistic regression which found six variables to be risk factors for mortality. This model's performance was significantly inferior possibly due to the neural networks ability to find non-linear relationships.

Prediction failures in ANN models indicate that information provided by a CDS system is not sufficient to decide not to treat or use as "an individual no-treatment policy" because it could be inaccurate for certain patients (like strong but small patients) [Zernikow et al 1998]. With a CDS system the clinician can "depersonalise his/her experience, and to make it accessible to junior colleagues" [Zernikow et al 1998].

There are few successful outcome prediction models for adverse events in the NICU, partly due to the failure of linear systems and the difficulty in classifying rare outcomes with high sensitivity and specificity.

In a study of mortality estimates, it was found that physicians predicted mortality with 90% specificity and 68% sensitivity, which means out of every 10 predicted survivals, there will be one death (false positive) and out of every 3 predicted deaths, there will be one survival (false negative) [Stevens et al. 1994]. Similar numbers were reported by Qi [2005] (specificity of 92% and sensitivity of 50%). How well these subjective judgements are founded have necessary implications for "patient triage or transfer, initiation and/or escalation of therapy, termination of life support, and allocation of medical resources", which outlines the value of well designed CDS systems [Stevens et al. 1994].

MIRG is working to develop models using a combination of ANNs, case-based reasoners and thermal images that can provide predictions for individuals. Past MIRG students like Khan [2006] developed a temporal model of mortality that incorporates day

3 data. Her model outlined thirteen important factors and the model classified mortality with 93 % specificity and 38 % sensitivity.

Hybrid ANN-CBR Imputation

Ennett [2003] was the first person to use linear (MSE) ANN weights in a CBR to impute values in a database. Her work provided a model of mortality that performed better than scoring systems alone. In its original format, the CNN database comprises primarily of incomplete cases, fragmented over multiple tables. Ennett, who first imputed the database, applied her novel weighted CBR method to fill the missing values. This was a critical step in Ennett's development of a new mortality model because considering only complete cases would have left too few cases for a generalizable model [Ennett 2003].

Ennett's imputation method consisted of a vertical imputation to build a base of 5102 cases with complete values for the nine SNAPPE-II variables, followed with a 3-step horizontal expansion to fill in the other fields. The vertical imputation was conducted with uniform weights in the CBR, where the mean of the 10 closest cases was used to replace the missing value. Cases with fewest missing values were imputed first, and then added to the match set for the following imputation to provide the CBR with the greatest quantity of complete cases from which to choose.

Prior to each horizontal expansion, a linear ANN model of mortality was created with the complete variables, and the relative weights of the ANN model were used to modify the distance metric in the CBR that was used to find the closest matching cases and fill in incomplete cases. Ennett justifies using relative ANN weights in the CBR by stating that "the weights in a neural network are presumed to reflect the importance that the network places on each input variable with respect to its **impact on the outcome under investigation**". Therefore we reasonably expect that the matching weights of the CBR would be the same [Ennett 2003].

The horizontal expansion is illustrated in Figure 3.6. The weights of the ANN model created with 9 variables were used as weights to impute 949 cases to fill the 5102 cases to 19 variables. The weights of the ANN model with these 19 variables were used to impute 2194 cases, to expand the database to 24 variables, and the process was

repeated again to obtain complete values for the 30 variables in all 5102 cases. Each step had a distribution of ANN and CBR weights ranging between 1 and 100 [Ennett 2003.

Once a database of 5102 cases with 30 complete variables was obtained, Ennett used the ANN to reduce the 30 variables and determine a minimal dataset for the

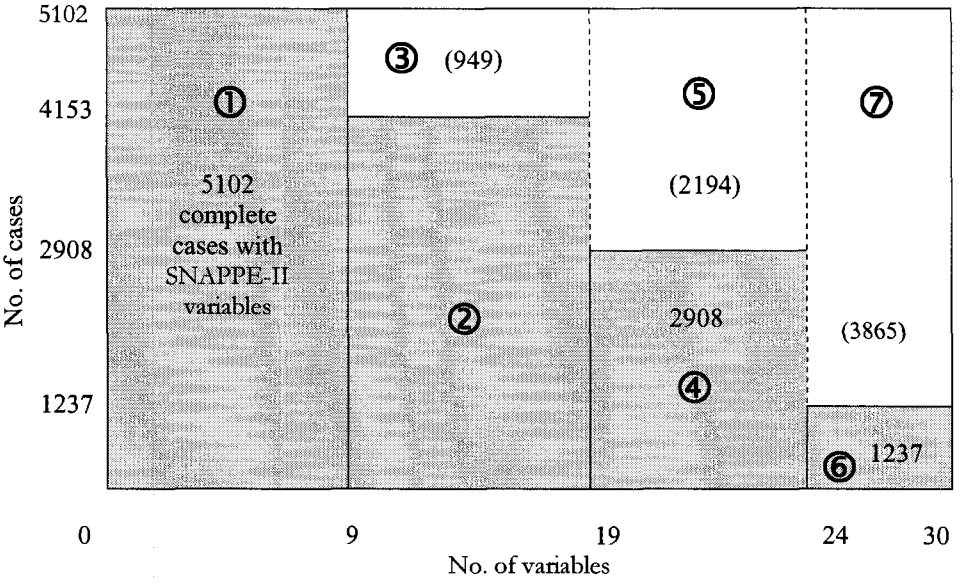

Figure 3.6. Step-wise horizontal expansion [Ennett 2003]

prediction of mortality comprising of 13 variables: *po2fio2r*, *Lurine*, *Lserum*, *apgar5*, *Lplt*, *SGA*, *Hsodium*, *Hrespr*, *Hpco2*, *bthwght*, *Lgluc*, *Ltempf* and *Hbloodp*.

Lastly, a vertical imputation was performed to fill in the 13 remaining values in the remaining fourteen thousand cases, again using the ANN weights as CBR weights and starting with cases having one missing value, then continuing through many imputation cycles until no missing values remain. This outcome-dependent imputation produced excellent results when classifying mortality, while all other outcomes that were modeled by subsequent MIRG students obtained promising but inferior results [Rybchynski 2005], [Qi 2005]. One of this thesis' goals is to determine if the reason for these inferior results can be attributed to the bias and clustering of cases created by

unequal ANN weights, or because some important variables necessary for the prediction of other outcomes were not retained in the database.

Ennett verified the performance of a model for mortality that used uniform weights in the k-NN, but only with the 13 variables that had been found using the outcome dependent weights. Results showed that the ANN-CBR variables were highest performing on ANN-CBR imputation.

Table 3-3. ANN Results (test set mortality rate = 3.4%, n = 19427) [Ennett 2003]

Performance measure	ANN-CBR	Clinician weights	Uniform weights	Mean	Random values
Sensitivity (%)	44.5	46.3	40.1	35.2	35.2
Specificity (%)	97.7	97.4	98.3	98.6	98.7
CCR (%)	95.9	95.6	96.3	96.5	96.5
ROC	0.8659	0.8561	0.8627	0.8296	0.8587

Though the results of the ANN using uniform weights for the final imputation are very close to that of the model obtained with the ANN weights, it is possible that the imputation method would affect the minimal dataset it if were used throughout the database expansion and reduction process.

Zernikow et al. [1998] developed a model specific to preterm neonates which achieved a higher performance that the hybrid model using a database of 890 babies. "A possible reason that Zernikow et al. chose this patient population was because there were no missing values in the database. This obstacle was challenged with the development of the ANN-CBR hybrid system's model by imputing missing values using the mean of the closest-matching cases" [Ennett 2003].

3.3.2 Length of Stay Models

Zernikow et al. [1999] also used ANNs to predict the length of stay in the NICU (**LOS**) of preterm babies using admission day data. Using 2144 cases, which were selected because they contained complete data for 40 CRIB, SNAP and maternal variables, a multiple linear regression model was created to extract the 14 most significant variables (only 5 of which are available in CNN). The performance of the

ANN model was measured in terms of correlation between actual and predicted LOS, rather than sensitivity and specificity as in their previous study. The ANN model performed better than the multiple regression model. The authors mention that having all cases abstracted from the same hospital resulted in a uniform discharge policy for all the cases and likely had increased the prediction accuracy. As well, using complete cases only may have helped classification results.

Interestingly, the researchers detected an inverse correlation between the LOS and the time elapsed between the baby's birth and the study date, which could suggest a recent trend towards more liberal discharge criteria [Zernikow et al. 1999]. Some variables, such as physician's responses to illness severity could have influenced the results if practices changed towards a more aggressive treatment policy. In general, the assumption made for the use of ANNs is that "the processes which give rise to the data do not themselves evolve with time" [Bishop 1995, p. ix]. The study used files from patients who were hospitalized over an eight year period (1989-1997), and it could be hypothesized that the selected charts provided a non-stationary source of data, which were difficult for the statistical approaches and ANNs to model.

Using the database that was imputed with the mortality weighted k-NN, Rybchynski studied the possibility of predicting other outcomes: length of stay and duration of ventilation. However the difference in relative importance of the variables for each outcome lead Rybchynski to conclude that "a minimum dataset must be created for each individual outcome. Subsequently, prediction models should not be used to predict anything other than their primarily intended outcome" [Rybchynski 2005].

Table 3-4. Relative importance of SNAPPE-II variables for three outcomes

	Bthwght	SGA	Apgar5	PO2flO2r	Lurine	Ltempf	Lserum	Lbloodp	Seizure
Mort	58	51	96	100	67	59	56	65	54
Vent	100	26	35	82	39	23	51	43	38
LOS	100	24	38	32	46	33	39	32	33

These results and conclusions also support an effort to impute a database in an outcome-independent fashion in order to use it to predict multiple outcomes.

3.3.3 Performance of Other Outcome Prediction Models

MIRG is working at finding models for adverse events in clinical settings with the aim of developing high quality screening tools and CDS systems. Preliminary research on the predictability of complications using admission data revealed that complications can be modeled, though not as satisfactorily as mortality using the 9 SNAPPE-II variables and other complications and procedures as input variables [Frize & Zhou 2006]. These encouraging results motivated the work into finding improved models using a wider number of more appropriate variables.

CHAPTER 4 CRITICAL ANALYSIS

The accuracy of physician's predictions for NICU outcomes "has obvious implications in terms of patient triage or transfer, initiation and/or escalation of therapy, termination of life support, and allocation of medical resources" [Stevens et al. 1994]. The predictions also play an important role in parent counselling. Very few studies have documented the precision in physician's risk estimates, in part because it is costly, time consuming, and there is a lack of individual outcome prediction models to serve as a comparison.

This research involves the development of prediction models and case presentation software as well as the validation of these models with physicians, which are important steps in the work towards clinical implementation of MIRG's CDS system.

4.1 Models Obtained with the Hybrid Imputation

Ennett's [2003] hybrid (MSE) ANN-CBR imputation of missing values in the CNN database began with a nine variable mortality model. Their relative weights were used in a CBR to expand the database horizontally to include more variables. The weights, which ranged from 1 to 100, reflected the relative importance of complete variables in the classification of mortalities. This step was repeated three times until all cases in the database contained complete values for the 30 SNAP variables. Table 4-1 lists the weights used in the k-NN for each horizontal expansion. The variable names are defined later in table 5-1.

The missing values were imputed with outcome-dependent weights and this database produced excellent classification results for the mortality outcome. Since Ennett's minimum dataset for estimating mortality produced a model with improved results over the best known scoring systems, the focus of MIRG's research work with this database evolved from mortality to include other outcomes and complications. While it is likely that the minimal variable sets of other outcomes would contain variables in common with Ennett's mortality model, it would be incorrect to assume they would all have the same relative importance [Rybchynski 2005].

Table 4-1. Weights used in k-NN [Ennett 2003]

SNAPPE-II		Set A		Set B		Set C	
Variables	Weights	Variables	Weights	Variables	Weights	Variables	Weights
bthwght	**100**	**Lurine**	**100**	**po2fio2r**	**100**	**Lurine**	**100**
apgar5	**62**	**po2fio2r**	**99**	**Lurine**	**75**	guaiac	92
seizure	**43**	guaiac	90	Hrespr	74	Hrespr	78
Ltempf	**41**	Hhema	85	guaiac	63	plots	53
Lserum	**25**	Hrespr	69	**apgar5**	**60**	Lplt	47
po2fio2r	**20**	**SGA**	**62**	Lplt	48	**SGA**	**47**
SGA	**15**	**Lserum**	**44**	**SGA**	**34**	Hgluc	42
Lurine	**3**	**Ltempf**	**37**	hpco2	33	Lhema	41
Lbloodp	**1**	Hheartr	33	**Lserum**	**32**	photos	30
		apgar5	**33**	**bthwght**	**31**	**apgar5**	**21**
		bthwght	**25**	Hgluc	29	**Lbloodp**	**19**
		Lbloodp	**24**	**Ltempf**	**29**	Hhema	17
		apnea	16	Hitnrat	29	Lgluc	16
		Hbloodp	11	Lgluc	25	**bthwght**	**16**
		Lheartr	11	Lhema	23	Hsodium	15
		Lgluc	10	Hheartr	20	Hitnrat	14
		Hgluc	6	**Lbloodp**	**18**	Lwbc	14
		Seizure	**2**	Lheartr	16	Lanc	13
		lwbc	1	Lanc	13	Hbloodp	11
				Lwbc	12	Lheartr	11
				Hbloodp	10	Hheartr	10
				apnea	8	OI	9
				Seizure	**4**	apnea	8
				Hhema	4	**Ltempf**	**6**
						Seizure	**6**
						Lsodium	3
						Lpo2po2	3
						Lserum	**2**
						Po2fio2r	**1**
						Hpco2	1

* Variables in bold belong to SNAPPE-II.

It was shown that the SNAPPE-II variables are ranked differently when outcomes other than mortality are modeled, which was expected because there are numerous contributing factors for each illness or diagnosis considered, which may be reflected in the importance of the collected physiologic parameters. Since the order of importance of the variables changes with the outcome modeled, it can be reasoned that variables imputed following Ennett's novel methodology would vary with the outcome used to supply the weights for the k-NN.

The database with imputed values based on the mortality model was used to model other outcomes; however results were not entirely satisfactory. Table 4-2 lists the relative importance of the SNAPPE-II variables when they were used to model mortality, extended length of stay (≥8 days) and short term ventilation (≤4 hours).

Table 4-2. Importance of SNAPPE-II variables in multiple outcomes [Rybchynski 2005]

	Bthwgt	SGA	Apgar5	Po2flo2r	Lurine	Ltempf	Lserum	Lbloodp	Seizure
Mort	6	9	2	1	3	5	7	4	8
LOS ≥8	1	9	4	7	2	5	3	7	5
Vent ≤4	1	8	7	2	5	9	3	4	6

A committee of classifiers was later constructed to improve on the ANN results. Satisfactory results were obtained with the mortality model, however the performance of other outcomes, especially in terms of specificity, remained inferior as seen in table 4-3. The CCR of the LOS model dropped well below that of the constant predictor, and while sensitivity remained high in the ventilation model, specificity fell to 50% from 85.9%. To obtain acceptable results in the real-time system, it needs to be determined if missing values can be imputed once with uniform weights in the similarity equation of the k-NN or if each model requires its own outcome-specific imputation.

Table 4-3. Results from the committee of classifiers [Rybchynski 2005]

	Mort	LOS ≥8	Vent ≤4
Sensitivity (%)	62.5	26.3	82.2
Specificity (%)	94.1	9.1	50
CCR (%)	89.8	25	74.6
CP	86.4	75	76.3

This research re-imputes missing values in the CNN using outcome independent (or uniform) weights. The performance of the models obtained help to determine if, in a real-time system, a database should store one replaced value per missing value or a list containing one replaced value per outcome predicted by the system. Resolving this question will affect the design of the data storing and retrieval architecture.

The following sections present a critical analysis of a few aspects of the initial prediction study, focusing on the ANN models, the interface and the questionnaires. The last section serves to summarize important points and suggest additional improvements.

4.2 Summary of Initial Pilot Survey

In 2005, a pilot study was conducted by a MIRG student [Qi 2005] to gather qualitative information on neonatologist's predictions for the outcomes of mortality, length of stay beyond 28 days and duration of ventilation exceeding 24 hours. Doctors' attitudes and impressions towards a software tool (ANN models and interface) were also collected. Overall results were slightly optimistic, though some physicians stated significant concerns about the predictive ability and usability of the tool. The study revealed many interesting observations and somewhat positive prediction results and attitudes. Concerns with the construction of the models and the presentation of the cases were also outlined and these were addressed before this new study took place.

4.3 Critical Analysis of Survey

The initial prediction study conducted in 2005 was encumbered by some difficulties, which this thesis explored and aimed to rectify. In 2005, Qi created models for mortality, length of stay and duration of ventilation with the nine SNAPPE-II variables. An acceptable mortality model was created because the SNAPPE-II variables were known predictors; however the models for the two other outcomes achieved inferior results [Rybchynski 2005], [Qi 2005].

MIRG's mortality model developed by Ennett [2003] had been shown to provide an improvement over the nine variable model used in Qi's study but was not used because EPIC does not contain 6 of the required 13 variables and Qi elected not to impute entire variables.

Qi used nine variables from the CNN database with Ennett's imputed values to build MSE ANN models (19398 complete cases with 3.74% mortality). The methodology stated that the "725 cases with outcome of death were extracted from the database and duplicated to create 1450 cases. These cases were added to the database,

which then contained a total of 20848 cases" and a 10.4% mortality rate to help improve training time [Qi 2005]. This new larger database, which contained each mortality case in triplicate was divided in to a training set (2/3 or 13899 cases) and a test set (1/3 or 6949 cases) and presented the classification results shown in table 4-4. The correct classification rate is above the constant predictor in all categories. In the case of mortality though, the CP is that of the CNN with the non-survivors in triplicate, not the true CP which should be 96.26%.

Table 4-4. Classification performance of the MSE ANN on the test set [Qi 2005]

	Mort	LOS ≥8	Vent ≤4	Vent ≤4 (linear)
Sensitivity (%)	70.2	71.0	76.7	63.4
Specificity (%)	93.0	85.9	90.6	89.0
CCR (%)	90.6	83.4	84.8	78.8
CP (%)	89.5	83.1	58.8	58.5
ROC	0.88	0.84	0.90	0.82

These results may be favourably biased towards correct predictions because validation sets were not used, therefore the generalization ability of these models could not truly be assessed. The fact that both training and testing sets contained 10.4% mortality cases implies that some mortality cases were present in both sets, causing the deaths in the test set NOT to be unseen data. This also could have given artificially improved classification results. Lastly, the test set did not accurately reflect actual distributions of outcomes which make it difficult to rely on the results of table 4-4 for an accurate assessment.

Imputing missing values from the EPIC database was not a significant task for the 2005 study because only nine variables were considered. Missing values (97 in total) were replaced by normal values under the assumption that they were indeed normal and would have been measured if they had been important for clinical management [Qi 2005]. As stated in section 2.4, it is incorrect to assume all missing values were normal since it was established by Richardson et al. [2001] that time or equipment limitations and/or pressing medical concerns may have caused a value to go unrecorded. If we accept the hypothesis that imputing missing values in the CNN with a weighted k-NN can

affect calculation of the similarity measure and therefore the replaced values, then missing values in EPIC should have been imputed using the same method.

The ANN models derived from the CNN cases showed inferior results when applied to the EPIC database; perhaps because overtraining could not be detected. In fact, degradation in performance in almost every category was observed. It is difficult to discern if this occurred due to differences in databases (1996-97 vs. 2003), or because EPIC has so few cases that the results are susceptible to the influence of individual cases. Questions can be raised about the generalization ability of the models, differences in variables between the databases and the effect of using a different imputation method with EPIC. It is important to note however that the inputs were limited to SNAPPE-II variables as opposed to NTISS variables which have been shown to be influenced by changes in practice over time as per Zernikow et al. [1999]. This may have helped reduce the influence of the time lapse between the compilation of the two databases.

The pilot study's prediction results are compiled in two tables to display the classification abilities of physicians, and to allow for a closer inspection of the results. Table 4-5 indicates that in comparison to the ANN models, physicians had higher sensitivities in predicting mortality, equal results for LOS sensitivity and somewhat inferior results for ventilation outcomes.

Table 4-5. Classification of the EPIC database by the tool and physicians [Qi 2005]

	Mort		LOS ≥8		Vent ≤4		
	ANN	physicians	ANN	physicians	ANN 3-layer	linear	physicians
Sens (%)	25	**50**	63.2	63.2	**87**	40	**73.3**
Spec (%)	98	92	63.6	36.4	0.0	100	78.6
CCR (%)	88.1	86.4	63.3	53.3	66.1	54.2	74.6
CP (%)	86.4	86.4	63.3	63.3	76.3	76.3	76.3

The study results in table 4-6, show that physicians correctly identified 4 out of the 8 deaths in EPIC, while the ANN model identified only two. The physician's strengths were classifying ventilation duration and long stays.

Table 4-6. Actual, physician and ANN classification of EPIC [Qi 2005]

	Mortality		LOS		Vent (2-layer)	
	Death	Survival	>28d	≤28d	≥24h	<24h
Actual outcome	8	51	19	11	45	14
Correct Predictions						
Physician	4	47	12	4	33	11
ANN	2	**50**	**12**	**7**	18	**14**

Their specificity was very high for mortality, less for ventilation and one third for LOS, which means that nearly 2 out of every 3 long term stay predictions resulted in a false positive. The strengths of the ANN were correctly classifying survivors, identifying short term ventilation and in classifying both lengths of stay. The ANNs' CCR was above or equal to the CP in two cases (Mortality and LOS) and below for ventilation models.

Due to the small size of the EPIC database and its few mortality cases (8), the results may not be representative of a larger study and are difficult to extrapolate to other outcomes or to other physicians. However, the results do demonstrate that the tool had some problems with generalization and some areas of success, as did physicians. Interestingly, the physician's prediction accuracy for mortality was consistent with the numbers reported by Stevens et al. [1994] which used eighteen physicians.

4.3.1 Case Presentation and Prediction Interface

In order to obtain outcome predictions for the EPIC database from neonatologists participating in the survey, Qi constructed scenarios for each case. This significant and time consuming task involved annexing a maternal history to each EPIC case to provide doctors with a more complete picture. These 40 variables contain sufficient information for doctors to assess a case and estimate outcomes. The variables, listed in appendix A table A-4, are salient information for five information categories: admission information, SNAP variables, mother's information, NITSS variables and any diagnoses or procedures that occurred within the first 12 hours of admission.

Upon review of the study questionnaires, some problems with the case presentation software were detected. It is believed that some of these problems were directly responsible for reduced physician acceptance of the ANN and lessened physician comprehension and acknowledgement of the predictive abilities of the ANN models.

In the construction of the tool for the initial prediction and attitude test, important issues were neglected which resulted in a diminished performance of the ANN models with the EPIC data and decreased clinical acceptance of the tool. Problems with the interface's clarity and the way in which the ANN predictions were presented to doctors prevented them from properly assessing the tool's usability and predictive ability. The critical approach of doctors and their traditional reluctance to adopt new technologies was overlooked. To illustrate the ANN's predictions, the interface of Figure 4.1 was built by Qi and used to display the case data and predictions to doctors.

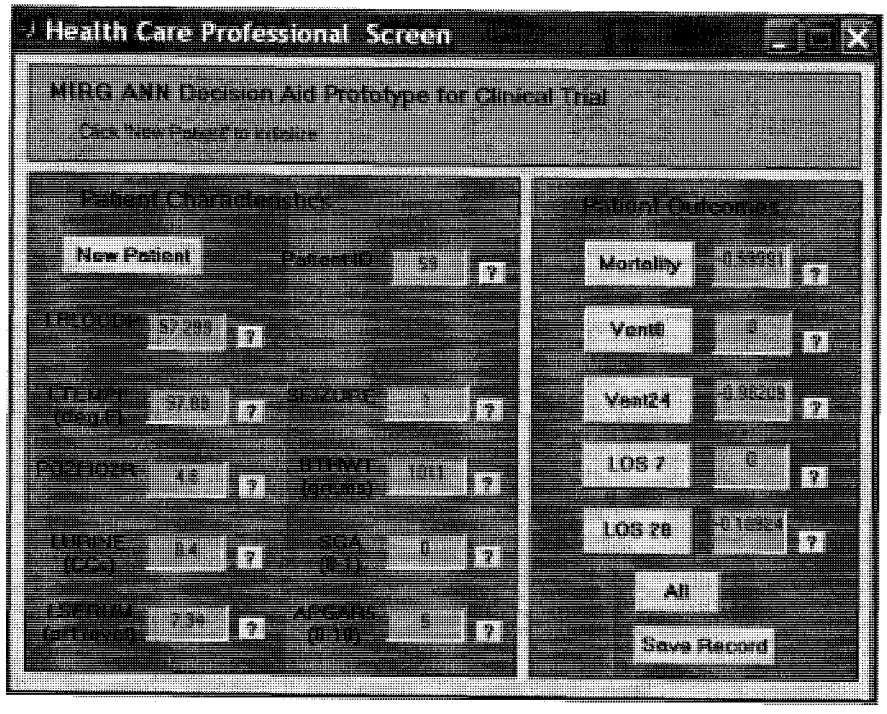

Figure 4.1. Interface provided to doctors [Qi 2005]

Qi presented doctors with the Matlab interface which displayed only the 9 input variables of the ANN models, as seen in Figure 4.1, when asking doctors if they agreed with the predictions of the ANN. Later, Qi concluded that the doctor's cognitive process

requires them to have the more complete list of variables in table A-4 to assess the prediction made by the tool [Qi 2005]. The completed questionnaires revealed that the manner in which the tool's predictions were displayed failed to gain doctors' confidence. Upon further analysis, the display can be found problematic from three points of views:

Mathematical: The ANN was designed to provide a binary response: the predicted outcome is mortality OR survival; not a percentage of risk situated between the two outcomes because a classifying ANN was used as opposed to the maximum likelihood estimation method described in section 3.1.2. Numbers in the 'Patient Outcomes' panel were therefore inexact. All negative numbers should have been rounded to -1 and positive numbers to 1.

Usability: 'Patient Outcomes' were presented as a continuous value between [-1, 1], without a written indication of what the numbers represented. The numbers could be understood by developers and expert users of the ANN software, but were not intuitive for new users. From a usability perspective, this format failed to be an efficient and an effective display of the predictions.

Another reason for the decreased usability of the interface is that the names of buttons were not always helpful. For example it is not intuitive to all users that "Vent24" represents "patient will require ventilation for more than 24 hours" and that "LOS 7" means "the patient will be hospitalized for less than 7 days".

The Lurine variable should be presented in the format that doctors are used to, in cc/kg/hr, as defined by the SNAP collection protocol as a *CCs* value does not correlate with doctors expectation of clinical usage [Dr. Bariciak 2006]. Lastly, the Seizure variable is coded in the CNN and EPIC databases as a nominal variable. Values of 0, 1 and 2 represent no seizures, 1 seizure event witnessed and 2 or more seizure events witnessed. The display in Figure 4.1 may imply that is was collected as a continuous variable.

Appropriateness: One doctor commented that "the cases provided here are not rare and a bit hard to understand at times. I would not see great value in ANN for regular premature

infants having regular courses of disease..." which speaks to the poor usability of the GUI and again questions the suitability of the EPIC database for this particular application. It is possible that the lack of variables presented had the effect of simplifying cases or obscuring complexities. Since doctors had insufficient information to make a prediction themselves, it was difficult for them to judge whether they agreed with the prediction displayed on the interface. The interface was redesigned for this study to display the variables and outcomes in a format that is more familiar and intuitive to doctors.

In this thesis, the approach is to present the tool in such a way that its predicted outcomes may be fully considered by doctors; the prediction acquiring and prediction showcasing screens were identical. In other words, all available variables were presented to doctors when they where asked for their risk estimates and when they were asked to consider the risk estimates provided by the tool. In future trials, doctors should always be presented with the complete case information when they are asked to review a CDS system's predictions or risk estimates.

4.3.2 Conclusions from Questionnaires

When doctors were asked to make their predictions known, they were provided with up to 40 important variables arranged in five groups in a spreadsheet. In a clinical setting, neonatologists are likely to encounter the *Neonatal Transport Log*, which could contain up to the 40 variables given for the prediction test (table 4-7).

For Qi's 2005 study, the EPIC database was divided into 6 groups, each containing approximately the same proportion of each outcome [Qi 2005]. Each physician was given 10 sheets of paper with the variables of table A-4 and asked to predict whether each of the three outcomes had occurred (Yes/No). This enabled each case in the database to be predicted by doctors. The ANN predictions were not included in this step. It was decided to present the doctors with the cases as paper-based surveys which conforms to the type of media format doctors currently use most often (i.e.: paper charts and paper lab results). It did not conform to the planned implementation of the real-time CDS system, but it minimized setup time and may have helped the

neonatologists complete the study more quickly. After consultations with MIRG's neonatologist partners, it was decided to perform this thesis's survey on paper as well to minimize the time participants would be required to be away from their patients and because the *Neonatal Transport Log*, on which the interface is modeled, is provided to clinicians in paper only.

During the study more than one neonatologist stated that most of the cases were routine: not the kind of cases for which they would want decision support. Since the research work in this thesis has a larger scope, charts which meet specific criteria were abstracted for this purpose. In a questionnaire Qi asked physicians to rank the importance of variables when predicting outcomes and obtained the results of table 4-7.

Table 4-7. Ranking of variables by each physician [Qi 2005]

	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5
Physician 1	Birth weight	Po2fio2r	Lserum	SGA	**Gestational age**
Physician 2	**Gestational age**	Birth weight	Apgar5	Po2fio2r	Lserum
Physician 3	Birth weight	**Gestational age**	SGA	Ltempf	Lbloodp
Physician 4	Birth weight	Specific diagnoses	Ventilation type/needs	Po2fio2r	Apgar5
Physician 5	**Gestational age**	Birth weight	IVH grade 3/4	Congenital anomalies	Lbloodp
Physician 6	Birth weight, SGA, Lserum, Seizure	Apgar5, Lbloodp Ltempf, Hrespr Po2fio2r, Hpco2 Lgluc, Lplt	Hbloodp, Hsodium, Lurine		

Qi suggested that gestational age was missing from the SNAPPE-II variable list because it was listed prominently by three physicians but was not included in the ANN model. However, the mortality model is trained with birth weight and SGA as input variables and the gestational age variable would be confounding information. It is strongly correlated with birth weight and SGA, because the SGA status is determined using those two variables. Qi's conclusion may have originated from the knowledge that doctors use this variable in their cognitive model even though they could derive it, or at least a range, from the two mentioned variables. Gestational age is therefore an important (but not mandatory) variable for doctors, and its inclusion in the SNAPPE-II model would not necessarily improve classification results. Gestational age will

therefore remain included in the list of variables displayed to doctors, but will not be added as an input to the outcome models.

Doctors suggested adding future outcomes like the chance of a negative neurodevelopmental outcome and severe IVH, which are some of the greatest concerns with very and extremely low birth weight babies (<1500 and <1000 grams respectively) [Qi 2005]. Such predicted outcomes could help physicians and parents make decisions about aggressive and life supporting therapies. The grade of IVH is a variable collected in the CNN and is modeled in the new survey. Unfortunately there are very few longitudinal studies collecting detailed data about the relationship between admission day variables and chronic morbidity. Developing a model for neurodevelopmental difficulties is not currently feasible however it is of interest to MIRG for further investigation when sufficient data becomes available.

4.3.3 Attitude Test Questionnaire

In the second part of Qi's study, attitudes of physicians towards the ANN models were acquired by means of a questionnaire. Answers revealed that important issues had been overlooked during the design of the study, which limited the doctor's ability to understand and evaluate the tool's predictions.

When introducing new technology to any group of highly skilled medical professionals, it is crucial to address the clinical reluctance which is known to exist to a change in protocol. Questionnaires provide a great opportunity to obtain information about such issues. In Qi's study, doctors were not given any information to describe the prior validations of the tool that are described in both medical and engineering journals [Frize et al. 1998], [Ennett et al. 2001], [Walker & Frize 2004]. Providing evaluators with information about ANNs and some of MIRG's publications to demonstrate the medical and mathematical soundness of the models could have alleviated some of the suspicions doctors exhibited.

The questionnaire asked important questions about the evaluator's background. However supplementary questions could instead have been added to the section which measured physicians' attitudes. None of the questions were directly aimed at detecting reluctance to use the tool, particular concerns about the correctness of the predicted

outcomes, or ask if any information was lacking for physicians to pronounce their judgement about the tool. One doctor remarked that the tool could replace a colleague's opinion [Qi 2005], but not their own, which indicates a certain level of both trust and reservation about the technology.

In conclusion, Qi [2005] was the first to bring MIRG's MSE ANN models into the clinical environment and revealed new and important details about doctors' prediction abilities and attitudes beyond those provided by studies such as Stevens et al. [1994]. Qi's study was useful as a pilot as it helped outline areas for improvements.

4.3.4 Risk Stratification

A model which classifies outcomes in a binary fashion allows for a comparison with other methods and models more easily than a model that produced multiple output categories. Authors such as Stevens [et al. 1994], Richardson [et al. 2001] and Zernikow [et al. 1998] produced models with a binary outcome. In a clinical setting however such a format may not be as helpful. When doctors consult with parents, they provide expected outcomes in a risk category, not in absolute numbers or in a yes/no category because there is always a certain level of uncertainty [Stevens et al. 1994], [Richardson et al. 2001]. This thesis aims to develop risk-stratified models in addition to classification models to concur with the intended clinical use of the system.

CHAPTER 5 METHODOLOGY

This chapter describes the methodology used in this research in preparation for the clinical survey and discusses the following topics in greater detail:

1) The Mean (k-NN) method was used to impute missing values in the CNN database. The imputation was done in a step-wise process in order to maximize the number of complete cases available for each imputation cycle. An artificial dataset was created using the original complete cases to measure the imputation error on known values.

2) The MSE ANN was used to create non-linear models for each outcome using the imputed subset of cases from the CNN (n = 13584). The modified Garson-Goh variable weighing method then ranked the variables in order of importance according to each outcome to allow for the determination of minimal variable sets.

3) The ML ANN was applied to each minimal variable set to create models of the estimated risk.

4) The prediction tool's interface was designed and the questionnaires were established in accordance with the goals of the study. The test cases and their predicted outcomes were integrated with the interface for the clinical survey.

5) The study protocol and informed consent were created and ethical approval was obtained for the clinical study.

5.1 Replacing Missing Values

The modeling methods used in this research require that all input cases be complete. The original imputation of missing values in the CNN database used a mortality-centered method to determine weights for the k-NN, but this thesis aims to create models for multiple outcomes. Since early modeling efforts of outcomes other than mortality with the first imputed database were not entirely satisfactory [Rybchynski 2005], [Qi 2005], the incomplete cases of the CNN were re-imputed with a uniformly weighted k-NN to give equal emphasis to each variable for all of the outcomes.

5.1.1 CNN Database

The CNN's abstractor's manual mandates the collection of over one hundred parameters for each patient on admission day (day 1). Ennett [2003] originally started the case selection process by joining multiple admission day SPSS[2] tables provided directly by the CNN into one SPSS file, Missing_All_Cases.sav. Very specialized variables and tests that were only performed on a few babies were not included. Only variables that are routinely collected for most babies were selected in order to reduce the number of missing variables per case and to develop models which would be potentially applicable to similar databases. The 30 SNAP variables that were abstracted on day 1 provided the starting point for Ennett's imputation so this thesis made use of the same variables. A second SPSS file, flat1.sav, was merged with the first: it contained some duplicate fields (caselink, mortality outcome and administrative fields) as well as values indicating the presence of BPD, the degree of IVH and NEC which were necessary for this research. The duplicate fields were compared after the merger to ensure a correct match. Some cases were then removed as per Ennett [2003] because missing variables rendered a case unusable: 443 cases with missing mortality/rare outcome information, 161 cases without a birth weight and 39 cases missing a small for gestational age value and gestational age.

For the 19814 cases that remained, outliers were removed or adjusted according to the guidelines provided by SNAP [Richardson et al. 2001], the CNN's Abstractors Manual [CNN 2004], and Ennetts' *Technical Report on the CNN* [Ennett 2003]. Table 5-1 lists the 30 SNAP variables along with the variable name and percentage of missing values per variable.

Table 5-1. Missing value statistics for the CNN (n = 19800)

Parameter description	Variable name	Percent missing
Small for gestational age	SGA	0.00
Birth weight	bthwght	0.00
Presence of apnea	apnea	0.02
Stool guaiac	guaiac	0.02
Presence of seizures	seizure	0.03

[2] Statistical Package for the Social Sciences version 14. SPSS Inc.

Highest heart rate	Hheartr	0.22
Lowest heart rate	Lheartr	0.26
Highest respiratory rate	Hrespr	0.36
Lowest temperature	Ltempf	0.41
Apgar score at 5 minutes	Apgar5	1.73
Highest glucose	Hgluc	8.72
Highest mean blood pressure	Hbloodp	12.70
Lowest glucose	Lgluc	14.51
Lowest mean blood pressure	Lbloodp	16.07
Lowest white blood cell count	LWBC	17.42
Highest hematocrit	Hhema	18.56
Lowest platelet count	Lplt	28.55
Lowest absolute neutrophil count	LANC	28.57
Lowest hematocrit	Lhema	33.53
Highest immature/total neutrophil ratio	Hitnrat	35.77
Lowest serum pH	Lserum	39.85
Highest pCO2	Hpco2	40.26
Lowest urine output	Lurine	52.68
Highest sodium	Hsodium	59.53
Highest potassium	Hpotass	62.65
Lowest pO2	Lpo2po2	63.14
Lowest pO2/FiO2 ratio	Po2fio2r	64.11
Lowest sodium	Lsodium	66.00
Lowest potassium	Lpotass	68.04
Highest oxygenation index	OI	72.76

Three variables were eliminated from the table above because they are not typically tested in the first 12 hours: **Stool guaiac**, **highest potassium** and **lowest potassium** [Bariciak 2006]. The eliminated variables had a missing percentage of 4%, 63% and 68% respectively. Dr. E. Bariciak, one of MIRG's neonatologist partners for this research, expressed that the lowest value of blood CO_2 could be of interest for outcomes such as NEC and BPD. Unfortunately this variable is not part of the SNAP list and was not abstracted for the CNN database.

The distribution of complete and incomplete cases for the 27 remaining variables was analyzed to provide information for the imputation process. The distribution of cases according to the number of missing entries from complete cases (27 variables) to cases containing only three variables is shown in table 5-2. As the number of missing values

reached 9 and above, the number of cases per missing variable category generally decreased. Few guidelines exist on determining the maximum percentage of variables to impute per case. Increasing the number of training cases does not necessarily lead to a better model. At some point there is a diminishing return on investment: the inherent error from imputation increases and fewer cases are being filled every iteration. Both the number of missing values per case and the percentage of missing entries for a certain variable will influence the imputation error. The effect of each imputation cycle on the prevalence of the rare outcomes was also considered. Ideally, the largest number of rare outcomes would be selected for imputation to increase the variety of cases in the training set and the generalizability of the models.

Table 5-2. Distribution of missing values in cases that were candidates for imputation

Number of missing variables	Number of cases	Frequency (%)	Number of missing variables	Number of cases	Frequency (%)
complete cases	1122	5.67	12	472	2.38
1	1277	6.45	13	322	1.63
2	1168	5.90	14	621	3.14
3	1892	9.56	15	228	1.15
4	1636	8.26	16	808	4.08
5	1343	6.78	17	89	0.45
6	2054	10.37	18	233	1.18
7	1237	6.25	19	28	0.14
8	1855	9.37	20	7	0.04
9	1154	5.83	21	4	0.02
10	1299	6.56	23	23	0.12
11	925	4.67	24	3	0.02
			Total	19800	100.00

Many stopping points for the imputation were considered and it was decided to proceed with eight imputation cycles because this resulted in a maximum number of 8 variables (or 30%) imputed per case. The resulting database is of adequate size, 13584 cases, providing more than 10 times the number of training cases than the number of input variables. As well, 661 mortality cases were included in this database, which represents 91% of all mortality cases available.

It was found that as the number of missing variables increased, the percentage of mortality and rare outcome cases decreased; this is likely because fewer seriously ill

cases were considered. This agrees with Ennett's [2003] finding that the mortality rate was higher in cases with the fewest missing values. Important variable statistics for the complete and incomplete cases retained for imputation are reported in table 5-3.

Table 5-3. Distribution of missing values in CNN (n = 13584)

Variable	Min	Max	Mean	St. Dev	# Missing	% Missing
SGA	0	1	0.04	0.20	0	0.00
birth weight	280	6320	2338.65	1042.19	0	0.00
apnea	1	4	1.17	0.56	0	0.00
seizure	1	3	1.05	0.29	0	0.00
Hheartr	60	327	161.84	18.00	1	0.01
Lheartr	0	197	126.73	15.52	3	0.02
Ltempf	68	113	97.43	1.21	3	0.02
Hrespr	0	180	58.61	29.06	18	0.13
Apgar5	0	10	7.83	1.74	33	0.24
Hbloodp	0	121	49.07	11.31	260	1.91
Hgluc	0.9	65.3	5.44	2.59	341	2.51
Lbloodp	0	100	36.71	9.54	392	2.89
LWBC	0.39	1442	15.17	22.37	483	3.56
Hhema	0	82	49.37	8.84	624	4.59
Lgluc	0	20.7	3.42	1.63	988	7.27
Lplt	0.22	3337	212.63	102.80	1594	11.73
LANC	0	686000	8599.17	12648.83	1888	13.90
Lserum	4.4	7.79	7.31	0.11	2535	18.66
Hpco2	4.44	172	46.80	14.06	2600	19.14
Lhema	0	80	48.45	9.23	2796	20.58
hitnrat	0.00	1	0.16	0.16	2999	22.08
Lurine	0	11.73	0.89	0.82	5471	40.28
Hsodium	106	182	136.20	5.06	6395	47.08
Lpo2po2	0	434	61.09	33.47	6476	47.67
Po2fio2r	0	12.42	1.90	1.27	6642	48.90
Lsodium	86	180	135.51	5.19	7392	54.42
OI	0	275	9.10	16.13	8297	61.08

From the above table, we can see that some variables occupy a very narrow range while others have a much broader range. The missing percentage for these variables varies between 0 and 61%. However more than half of the variables have fewer than 5 % missing values.

Artificial Imputation

Prior to imputing missing values, an artificial dataset was created to measure the accuracy of the selected imputing methods. To do so, known values were randomly removed from the collection of complete cases (n = 1122) until the distribution of missing values equalled that of the complete database, as shown in the last column of table 5-3. The imputation error on known values was calculated to provide information about the uniform weights approach and about three other common approaches:

1) Mean (k-NN): a k-NN algorithm with uniform weights (a weight of 100 for each of the 27 variables) is used to find the 10 most similar cases and the mean of a variable is used to fill the missing value.

2) Mean: a missing value is replaced with the mean of the remaining values.

3) Normal: a missing value is replaced with a random value from a normal distribution having a mean and standard deviation equal to the remaining values. This replacement method is performed in SPSS with the syntax of equation 5.1:

IF sysmis(variable) variable = rv.normal(mean, standard dev) (5.1)

4) Uniform: a missing value is replaced with random value from a uniform distribution having a range containing 95% of the remaining known values. The range for the uniform distribution was reduced to 95%, or 2 standard deviations to lessen the effect of occasional outliers. Also performed in SPSS, this method uses equation 5.2 to impute missing values:

IF sysmis(variable) variable = rv.uniform(min, max) (5.2)

The percent error between the imputed value and the original known value was calculated to give some measure of the error incurred during imputation for each of the four methods described. Equation 5.3 defines the percent error. Following the calculation of the percent error, the average error was calculated with equation 5.4:

$$\text{percent error} = \frac{|(\text{imputed} - \text{known})|}{\text{known}} * 100 \qquad (5.3)$$

$$\text{average error} = \frac{\sum \text{percent error}}{\#\text{cases with missing values}} \qquad (5.4)$$

Both equations 5.3 and 5.4 were used to assess which method minimized the difference between actual and imputed values. A two tailed pair wise *t*-test was applied to each set of actual and imputed values to determine which imputation method produced values which were not significantly different ($p>0.05$) to the true ones. In view of the restricted time and workload involved in developing each model, observing the effect of imputation was limited to the SNAPPE-II variables which had missing values. These variables are considered general indicators of illness and have varied ranges and missing value percentages as seen in table 5-3. It is believed that they provide sufficient information for the selection of an imputation method.

Actual Imputation

After calculating the imputation error on known values using the complete cases only, the Mean (k-NN) technique was applied to cases with unknown missing values to create a larger database from which models were developed.

The first CBR cycle used the complete cases ($n = 1122$) as a match set to impute values in cases missing only one value ($n = 1277$). This new collection of complete cases was then used to impute cases with two missing values ($n = 1168$). After eight imputation cycles, the filled CNN database contained 13584 complete cases. The distribution of the SNAPPE-II variables with missing values was again tracked throughout the imputation.

5.1.2 EPIC Database

The EPIC database was collected from a tertiary care center implementing guidelines developed to lower the incidence of severe lung disorders and hospital-acquired infections in at-risk babies [Lee 2006]. The distribution of missing values was quite different than that of the CNN. Three possible reasons are: it is a much smaller database, it was collected with a different purpose and the cases were abstracted from a single site. No blood gas reports or ventilator parameters were recorded, which is unfortunate because many values they contain are very important for outcome prediction [Frize & Walker 2000], [Catley et al. 2005], [Zernikow et al. 1998]. It is known that some babies were ventilated and had blood drawn because the relevant NTISS variables

are reported, however actual laboratory results were not entered as they were in the CNN database. Values for the 17 available SNAP variables were extracted from the 50 tables making up the database, using the Record Number field to match all the variables to their respective cases.

Cases in the EPIC database had between 10 and 14 missing values each, as per table 5-4. Each case was expanded horizontally using the CBR system to include all 27 SNAP variables imputed in the CNN.

Table 5-4. Distribution of incomplete cases in the EPIC database (n = 59)

Number of Missing Variables	Number of Cases	Frequency (%)
10	18	30.51
11	17	28.81
12	14	23.73
13	5	8.47
14	5	8.47

For this thesis, the imputation process for the EPIC database employed the Mean (k-NN) method as it proved most accurate in the artificial imputation tests. The imputation was identical to that of the CNN, with the difference that the original complete CNN cases (n = 1122) were used to fill the missing values for the 18 cases with 10 missing variables in the first imputation cycle. Those 18 imputed EPIC cases were then added to the matching database to impute the 17 cases with 11 missing variables and so on until all EPIC cases were complete. A large percentage of variables were imputed; however the imputation error was overlooked to allow the EPIC database to be used with the models developed in this thesis. The EPIC database provided interesting classification challenges due to its small size and the high number of low prevalence outcomes. Imputing missing values allowed for a comparison with previous models to observe differences in classification performance which may result from the new imputation methods and a wider selection of variables.

5.1.3 CHEO Database

A small group of cases were collected under the guidance of MIRG's neonatologist partner for the purposes of this thesis. To create this new database, variables were individually abstracted from archived charts at CHEO onto a custom paper form. Case abstracting is a very laborious process and requires a very conscientious abstractor. The selected cases were from patients admitted in 2004-2006. The variables were entered onto a spreadsheet, but required significant processing before missing values could be imputed as discussed below.

The equations required to calculate SNAP variables from ventilator parameters are provided in the CNN guidelines and the CNN abstractor's manual. The oxygenation index was calculated with equation 5.5 when the individual variables were recorded within an hour of each other:

$$\text{Oxygenation Index} = (FIO_2 * MAP)/P_aO_2 \qquad (5.5)$$

where FIO_2 is the inspired oxygen concentration, MAP is the mean airway pressure of the oxygen and P_aO_2 is the partial pressure of arterial O_2.

The total amount of urine voided until the first diaper change after 14 hours post-admission was converted to cc/kg/hr, temperatures collected were converted from Celsius to Fahrenheit and units of pressure were transferred from cmH_2O to mmHg to match the units described in the SNAP guidelines.

The lowest and highest blood pressure values were transcribed onto the patient charts in two formats: the standard systolic over diastolic fraction and a mean value which is also displayed on the blood pressure monitor. To conform to the SNAP guidelines, blood pressures recorded as a fraction were converted to a mean value with equation 5.6:

$$\text{mean pressure} = (2 * \text{systolic} + \text{diastolic})/3 \qquad (5.6)$$

Lastly the Po2fio2r variable was calculated using the two required values (PO_2 and fIO_2) if they were recorded within the same hour; otherwise the field was left blank. Variables such as lowest platelet count and highest sodium value were often missing because the specific blood or electrolyte tests had not been ordered within the first 12 hours after admission or were missing from the file. Often, values necessary for the OI

and Po2flo2r variables were not measured at the same time. Only values which are correlated in time can be used to calculate these fractions because they are meant to represent a patient's status at a precise point in time. Variables not in SNAP units were left as collected for the survey in order to conform to physician's expectations.

The length of stay was calculated using the admission and discharge dates, and the duration of ventilation was calculated by adding the duration of all ventilation periods recorded during the hospitalization. A ventilation of a few hours, but covering two different calendar dates, was counted as a single day. The list of final diagnoses was scanned for the presence of NEC and BPD and for the degree of IVH using the International Classification of Diseases codes version 10.

Table 5-5 shows that again in this database, the average number of missing variables per case was greater than the eight missing variable limit imposed in the CNN imputation.

Table 5-5. Distribution of incomplete cases in the CHEO database (n = 60)

Number of Missing Variables	Number of Cases	Frequency (%)
7	6	10.00
8	5	8.33
9	14	23.33
10	17	28.33
11	18	30.00

Missing values in the CHEO database were imputed with the Mean (k-NN) method according to the same protocol as for the EPIC database for consistency with the CNN database, beginning with the six cases missing only seven values.

5.1.4 Outcome Distribution

The CNN database is by far the largest of the three, and most closely represents actual rates of low-prevalence outcomes in the general Canadian neonatal population because they are averaged over 17 NICUs. The EPIC and CHEO databases are very

small and the cases were collected with very specific guidelines and different purposes. This is reflected in the prevalence of outcomes.

The outcomes modeled in this thesis are mortality, hospitalization lasting 28 days or more, a period of 8 or more days of ventilation, a final diagnosis of BPD, a final diagnosis of a severe IVH and the occurrence of NEC. The distribution of outcomes is listed in table 5-6.

Table 5-6. Distribution of outcomes in the CNN, EPIC and CHEO databases

	Mortality (%)	LOS ≥28 (%)	Vent ≥8 (%)	BPD (%)	IVH (%)	NEC (%)
CNN (n = 13584)	4.87	27.31	14.20	8.12	3.72	2.38[a]
EPIC (n = 59)	13.56	37.29	25.42	40.68	unknown	13.56
CHEO (n = 60)	18.33	35.00	45.00	3.33	5.00	11.67
CNN (n = 20488)[b]	3.79	20.92	8.74	7.30	4.90	1.50

[a] n = 13133 [Zamboni 2007], [b] includes all complete and incomplete cases of the CNN

The severe IVH prediction model was not applied to the EPIC database because there was no information in the EPIC tables stating whether any patient had been diagnosed as having some degree of IVH. The very high incidence of BPD, combined with a smaller relative increase of long-term ventilation reflects the goal of the EPIC database which is to develop evidence-based strategies for the treatment and reduction of chronic lung disease. These charts may reflect treatment policies which resulted in patients having reduced morbidity. These cases are expected to be particularly difficult to classify correctly as they represent a small subset of cases within a rare outcome.

The CHEO cases were selected to present doctors with a denser collection of rare outcomes and complications. This is reflected in table 5-6 where the percentages of complications associated with the CHEO database are higher than those reported for the CNN, which may reflect the characteristics of sicker babies.

The prevalence of outcomes and complications varies greatly between hospitals. Any random collection of 60 cases from the CNN is likely to have a different distribution as shown by Qi [2005] and Rybchynski [2005]. Abstracting cases that were distributed more similarly to the CNN was not attempted partly because of the comments provided by neonatologists in the exploratory survey. The present scenario, though likely to lead to inferior classification results, is a more realistic presentation of the potential use of the

tool in various clinical situations and will provide more valuable information. The tool must be validated on the cases for which it is most likely to be used, as opposed to hand-picked cases where the tool is known to perform well. The rational for conducting a study with a more challenging database so that it can provide a better assessment and a more robust test of classification ability. Using a database with fewer "normal" cases may provide additional insight into specific areas of strengths and weaknesses of the tool. The disparity in prevalence of the outcomes serves to further emphasize the importance of generalizability while developing models for this type of CDS system.

5.2 Determining the Minimal Variable Set

With a large collection of complete cases from the CNN available the MSE ANN was used to create models for each outcome. The normalized data was split into three sets: 1/3 was first set aside for the verification set. Of the remaining data, a 2/3 portion was used for training and the remaining 1/3 for testing. In all cases except length of stay, the positive cases of the training set were artificially resampled until they comprised 20% of cases to help expedite training. All positive and negative cases used in the verification set were first removed from the database and were not included in the training or testing sets so that they consist of entirely new and unseen data. The number and percentages of common and rare outcomes cases in the training, testing and validation sets is reported in chapter 6.

Initially, 13584 cases with 27 variables were used as input to the MSE ANN. In order to reasonably determine if a certain group of variables could provide an improved model, structures with zero to 2n+1 hidden nodes, where n is the number of input variables, were constructed. The modified Garson-Goh parameter selection algorithm was applied to the weights and bias values of the best structure to obtain a ranking of the input values in accordance to their importance in the classification. The variables with the least influence on the model were removed until the minimal variable set was obtained.

5.3 Creating Risk Models

Since doctors present possible outcomes to parents in risk categories rather than in absolute percentages, a risk stratification process was applied to the best performing model for each outcome. First, the ML ANN was used to create a model for each outcome using the minimal datasets derived in the previous step. Once the best performing structure was selected for each outcome, the next steps involved sorting the cases into three risk categories that are appropriately categorized for use in the NICU: high (100%-75%), moderate (74%-21%) and low risk (20%-0%) [Bariciak 2007], and applying the models to the EPIC and CHEO databases.

The data preparation step was less complicated than with the MSE ANN configuration because the likelihood function requires an a priori distribution in all three datasets. Therefore no artificial resampling was performed. To produce a valid ML model, three datasets were created: a set containing all the cases, a training set containing 2/3 of cases randomly selected and a test set created from the remaining 1/3 of cases. Inconsistent distributions result in the failure of development a good risk estimation model; therefore an effort was made to ensure consistent distributions. For each outcome, each of the 2n+1 structures were evaluated with the Hosmer-Lemeshow goodness of fit test to assess calibration and by the ROC value to assess discrimination ability. A p-value above 0.05 and ROC above 0.85 were set as the requirements for all three datasets. The distribution of rare and common outcome cases in the datasets is reported in chapter 6.

5.4 Survey Components

5.4.1 Interface Design

Qi's survey's prediction interface was abandoned and a new concept was designed with the goal of presenting the cases and prediction to physicians in a clear and intuitive way. Comments acquired from physicians during the previous pilot survey were considered for the new design. The information physicians require to make their predictions is found in the Neonatal Transport Log; a CHEO-specific sheet that contains all available data about patient and mother, collected by the medical transport team. The

Neonatal Transport Log is always encountered by neonatologists at CHEO because all babies are transported to this NICU. Knowing that neonatologists are already familiar with a variable layout, a digital version of a Neonatal Transport Log was created to display the cases and the predictions. Usability and chance of clinical acceptance should be increased compared to the initial survey interface by providing more information to physicians when they are asked to review the tool's predictions.

Variables and outcomes were presented in a clearer and more familiar format, where the decimals that indicated the 'Patient Outcomes' were replaced with "present" or a "☑" symbol and "not present" or a "☐" symbol. The salient variables (Apgar, presence of seizures, etc.) were placed more prominently because they affect the assessment of the other variables in the case [Bariciak 2007]. The ventilation status at 12 hours was indicated on the interface.

The new interface was designed in MS Access to facilitate integration of the database containing the cases and risk category estimate and because it provides a user-friendly and familiar look. It was also reviewed by two neonatologists to check for unclear phrasing, terminology and variable placement. During the study, participants reviewed the case information presented by the interface and indicated their predictions on the bottom right portion of the display which is pictured in Figure 5.1.

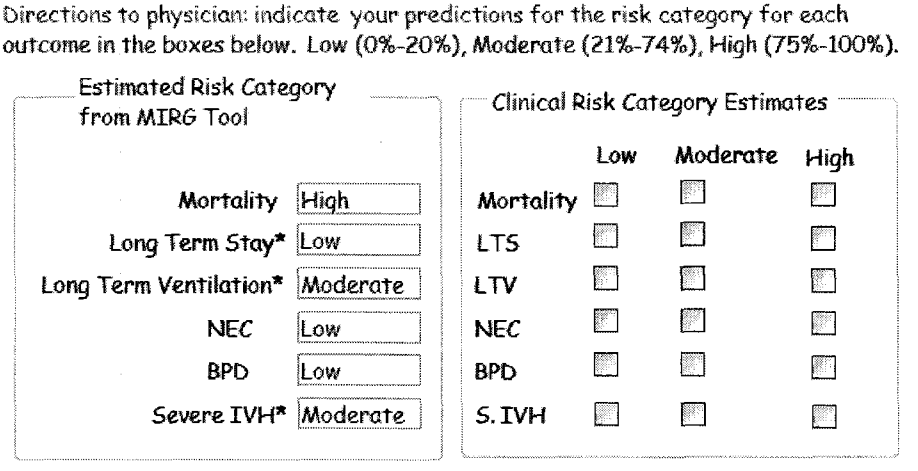

Figure 5.1. Area of interface for physicians to indicate predictions

For the first ten cases predicted to each participant, the risk estimates in the text boxes of the "Clinical Risk Category Estimates" were blank, and for the second set of ten cases the estimated risks were provided.

5.4.2 Questionnaire Design

The questionnaire in this study was designed to measure both the usability of the CDS system and the attitudes of physicians. The Technology Acceptance Model has been validated as a means to measure the intentions of use and perceived usefulness of an informatics system by physicians. The topics covered by the questionnaires were based on important principles described by this model.

The Technology Acceptance Model

The Technology Acceptance Model (**TAM**) evolved from the theory of reasoned action [Fishbein & Ajzen 1975], and is specific to informatics systems. The TAM uses two concepts to determine usage behaviour:

1) **Perceived usefulness,** "user's subjective probability that using a specific application system will increase his or her job performance within an organizational context" and

2) **Perceived ease of use,** "the degree to which a person believes that using a particular innovation would be free of effort" [Chismar & Wiley-Patton 2003].

The extended TAM model, Figure 5.2, aims to create a more comprehensive representation of information technology system acceptance and actual use.

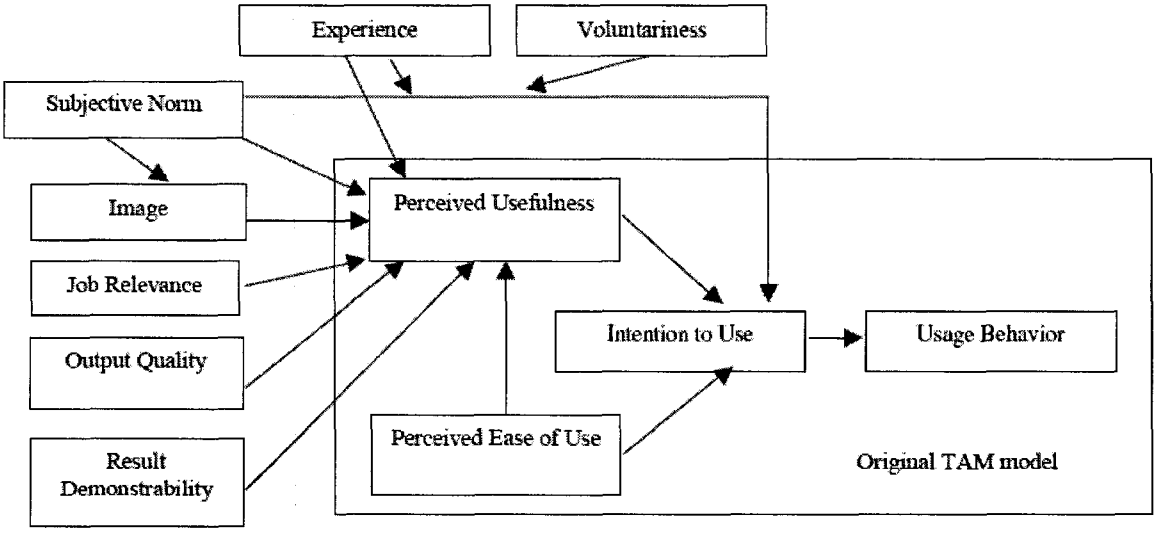

Figure 5.2. The extended TAM [Chismar & Wiley-Patton 2003]

At this stage in the development process, the perceived usefulness was especially relevant to physician's interest and desire to participate in future trials of the CDS

systems. The perceived ease of use will be improved when future trials are incorporated into the clinical workflow and the interface is finalized. MIRG is grateful for the continued collaboration and interest demonstrated by its medical partners.

The objective of this research was stated as follows in the Informed Consent:

> The overall goal of this research project is to compare the predictions of a proposed decision-support system for potential future clinical application using intelligent tools (e.g., artificial neural networks) with predictions of experienced neonatal physicians and observe the areas of success for each and determine areas that need improvement for the decision support system.
> A questionnaire records the attitudes of physicians towards the system.

This objective was read to potential study participants during the study explanation and repeated in the study protocol to transmit the goals and possible benefits of this CDS system.

Most healthcare institutions are moving towards a paperless environment and are increasingly relying on Internet-based applications to do so, such as the XML framework developed by Catley et al. [2005]. Many Internet-based health applications have been designed (electronic patient records, internet scheduling and prescription drug conflict alerts) but remain underused, which highlights the importance of proper usability studies and the collaboration of engineers, cognitive scientists, software developers and incorporation of user feedback during prototype review [Chismar & Wiley-Patton 2003].

The best way to measure the acceptance of a system with the extended TAM is with standardized, well-designed closed questions. Van Schaik et al. [2004] analyzed questionnaires based on the extended TAM which asked 30 general practitioners about Internet healthcare applications. They found that physicians' acceptance of an Internet application depended greatly on perceived benefits and was largely independent of their acceptance of computers.

In general, physicians were found to be much more pragmatic and would use a difficult application if it is useful. Perceived usefulness and output quality were found to be the most important constructs of the extended TAM by Chismar & Wiley-Patton [2003] who verified "the applicability of the extended TAM in the context of physicians' intention to adopt Internet-based health applications". This finding was echoed by their

2003 study of 89 paediatricians in which perceived usefulness was found to be a significant indicator of intention to use.

Providing predictions in risk categories was a strategy aimed at increasing the perceived usefulness as well as output quality by providing predicted outcomes in a more intuitive format and one that fits well with the manner in which physicians do their work.

This thesis did not measure voluntariness and experience in the questionnaires because physicians participating in the evaluation have never worked with such a system and are not facing the possibility of immediate implementation of this type of system. The extended TAM was also used as a template for questions relating to user acceptance of medical information systems. More specifically, the psychometric questionnaires developed by Ammenwerth et al. [2003] were used because they were shown to "measure what they are intended to measure (validity), and... do this in an objective and reliable way".

The closed questions in this survey employ a four point scale ranging from 'strongly agree' to 'strongly disagree'. The neutral or 'uncertain' option of the 5 point Likert scale was omitted to oblige the participants to make a polarized decision. By removing the neutral option, it is likely that apprehensive participants choose a less favourable option because physicians are known to be quite sceptical of protocol changes [Courtright et al. 2001]. This will make researchers aware of physicians' attitudes in a way that a neutral answer would not. The questionnaires used for this research are presented in Appendix D, and the topics covered can be summarized as follows:

The first questions focus on the case presentation interface and ask the participants about their feelings towards electronic tools in general and their current level of satisfaction with electronic tools designed for use in the NICU. Later questions also address the output quality in terms of display and outcomes presented.

Three questions cover the topic of likelihood to use the tool depending on whether the predictions agree or differ with those of the participant. Two questions directly address the perceived usefulness of the tool, while four questions address the participant's knowledge about ANN and CDS systems in general. Question 16 verifies whether or not participants prefer to have the outcome displayed in terms of risk. The

two last questions ask participants to list which outcomes and features, if any, they would like to see added in future versions of the tool.

The centre of attention is on system dependent features, assuming that doctors would be provided with all the software, hardware and technical support necessary to use the CDS system. Free text questions on positive and negative aspects of CDS were added to record "unexpected observations" such as perceived barriers. During the study, a research assistant was on hand because the questionnaires "do not allow for flexible interaction between researcher and user" Ammenwerth et al. [2003]. Extra lines were allocated on the questionnaires for participants to mention issues which they feel were not addressed in the questionnaire [Tang & Patel 1994]. These comments are important because they could allow researchers to detect aspects of the CDS system or clinical workflow which may have been overlooked during the development process and should be considered for future trials. For subsequent clinical trials and simulations, interactive interviews and more qualitative questions might be better suited to detect the unexpected.

The first and smaller section of the questionnaire gathered some demographic information about the survey participants such as the number of years of experience in their specialty and their feelings towards certain computer applications and uses. These questions will provide some information on the homogeneity and attitude of participants. The number of participants in the study was low because it is limited to one hospital. At this stage in research, major usability problems and development flaws can be detected with a small sample size, though it limits the possibility of quantitative and significant statistical analysis.

5.5 Study Protocol and Informed Consent

Establishing the study protocol and renewing ethical clearance from CHEO was an important step which helped clarify the objectives of the research. A copy of the Informed Consent and the questionnaire was provided to the Ethics Review Committee along with a detailed methodology. Because of this study's design and objectives, it is classified as low-risk and was approved under the rubric of expedited review. Qi's 2005 application was used as guide for the renewal, though the Informed Consent and questionnaires were modified in this new study.

Most studies bring about certain risks and benefits for the participants. When addressing risk in a study, the two principal options are to modify the purpose and protocol to eliminate all risk, or aim to reduce the risk as much as possible without compromising the goals of the study and then provide clear information for obtaining informed consent from the participant. The latter option was chosen in this research as it allows for a more detailed study and gives the participant the power to make his or her own decisions. Since this study uses retrospective case data without identifiers and the prediction results will be released as a group, no risks were foreseen for the participants.

The Informed Consent document is crucial in all studies as it protects and guides the participants. For this study, the Informed Consent describes the purpose, the procedure, risks and benefits of the study. The Informed Consent document also provides participants with a name and contact number should they have any questions regarding the study. According to the study protocol, participants must be informed verbally and by the Informed Consent that they have the right to stop participating in the study at any time and decide not to hand in their questionnaire without penalty or fear of reprisal.

Since the participants in this study can be considered a captive population (they are co-workers of MIRG's physician partners), measures were taken to ensure the study process was ethical for all involved and free of coercion. The most important ones are listed below:

1) Cases without risk estimations were presented to physicians first, to remove the possibility of physicians being influenced by the prevalence of rare outcomes signalled by the models.
2) Predictions made by the neonatologists were revealed in a group, not on a per physician basis, to avoid the fear of reprisals if a physician is found to have predicted less successfully than others have. Correct/Incorrect predictions will not be associated with particular cases in any publications or documents. Participants were advised of this protocol verbally and in writing in the Informed Consent.

3) The distribution of rare outcomes was as equal as possible between the physicians and between the cases presented with and without the risk estimates.
4) The study was explained to participants in its entirety prior to providing the Informed Consent.
5) The study was performed in a teaching hospital where the willingness to participate in studies is thought to be greater than in non-teaching hospitals.
6) Participants were not compensated regardless of whether they completed the study or not and there were no consequences for not participating.

Study Protocol Summary

The study protocol can be summarized as follows: potential participants were made aware of the study through an invitation to present at rounds by physicians at CHEO and posters in the NICU. Interested neonatologists presented to c designated person at CHEO on the date designated on the posters and were informed of the study and were handed an envelope with all necessary information for the completion of the study. Those who signed the informed consent made their predictions for 20 cases: the first 10 without risk estimates and the second group of 10 with risk estimates. Following the physicians' review of the 20 cases, he/she completed a questionnaire. Completed questionnaires were sealed in an envelope and did not contain any identifiable information other than the answers that were provided.

Section 6.4, *Integration of Components*, describes how the variables, risk estimates and acquired predictions were tracked and associated to the correct record number and actual outcomes for the data analysis.

It is in the interest of the participants, but also in the interest of the researchers to design a study protocol which respects the rights of all involved, including the right not to participate. It is believed that the protocol in this study complies with this requirement.

CHAPTER 6 RESULTS and DISCUSSION

This section discusses the imputation of the CNN database, the creation of MSE ANN models and estimated risk stratification. Later, the preparation work for the survey as well as the qualitative and quantitative survey results are presented.

6.1 Imputation Work

Error on Known Values

Objective 1 of this research was stated as follows in section 1.3:

> Confirm the ability of a case-based reasoner making use of a k-NN algorithm with uniform weights to impute missing values into a dataset by comparing imputed values to known values and estimating the imputation error.

In accordance with this objective, an artificial dataset was created by deleting known values from the original complete cases of the CNN (n = 1122). Four of the nine SNAPPE-II variables were complete in the CNN cases retained for imputation (birth weight, Apgar5, SGA and seizure) and the Ltempf variable had a very low missing value percentage (0.02%). The four remaining SNAPPE-II variables with missing values were analyzed. In the artificial dataset, the percentage of deleted entries for a variable corresponded to the actual percentage of missing values for that variable in the entire CNN database. The accuracy of the imputation methods was assessed by calculating the average percent error and observing changes in each variable's distribution caused by the imputed values. The percentage and number of missing values in the CNN database and artificial dataset is shown in table 6-1. The last row of the table determines the number of values removed from the artificial dataset.

Table 6-1. Missing SNAPPE-II values in the CNN database and artificial dataset

	Lbloodp	Lserum	Lurine	Po2fio2r
Missing % in CNN [a]	2.89% (n = 392)	18.66% (n = 2535)	40.28% (n = 5471)	48.90% (n = 6642)
Missing % in Artificial dataset [b]	2.85% (n = 32)	18.63% (n = 206)	40.29% (n = 452)	48.93% (n = 549)

[a] (n = 13584), [b] (n = 1122)

From the complete cases, 32 Lbloodp entries were removed. Next, the cases were randomly shuffled and 206 values from the Lserum column were removed. Again the cases were randomly mixed and 452 values of Lurine were removed and after re-mixing the cases, 549 values of Po2fio2r were deleted. In the artificial dataset 267 complete cases remained and 518, 293, 41 and 3 cases missing one to four values respectively. Three additional copies of this artificial dataset were created and the missing values in each set were imputed with a different method: Mean (k-NN), Mean value replacement, Normal Distribution replacement and Uniform Distribution replacement.

For the first pass with the Mean (k-NN) method, the 267 complete cases were used as a match set to impute one value in each of the 518 cases with 1 missing value. The new set of 785 complete cases (267+518) was used to impute values in the 293 cases with two missing values, and so on until all cases were complete.

Table 6-2 displays relevant information about the imputed variables in the artificial dataset of the CNN: the mean and standard deviation of all original known variables or 'True Values', the mean and standard deviation after the random deletions, and the range which contained 95% of the remaining values.

Table 6-2. Artificial dataset with actual, missing and imputed values

	Lbloodp	Lserum	Lurine	PO2fiO2r
True Values [a]	33.32±8.87	7.28±0.13	0.80±0.88	1.55±1.19
Remaining Values	33.25±8.84	7.28±0.14	0.83±0.85	1.58±1.21
95% Range	[12.724, 53.00]	[6.64, 7.47]	[0.00, 3.24]	[0.122, 4.62]
Imputed Values				
Mean (k-NN)	33.57±5.81	7.30±0.06	0.80±0.27	1.53±0.68
Mean	33.25±0.00	7.28±0.00	0.83±0.00	1.58±0.00
Normal	33.25±8.84	7.28±0.14	0.83±0.85	1.58±1.21
Uniform	33.39±12.40	6.99±0.21	1.68±0.89	2.31±1.30
No. of cases	**32**	**206**	**452**	**549**
Imputed Dataset (n = 1122)				
Mean (k-NN)	33.26±8.80	7.28±0.13	0.82±0.68	1.56±0.99
Mean	33.25±8.71	7.28±0.12	0.86±0.66	1.58±0.86
Normal	33.26±8.84	7.28±0.14	0.83±0.85	1.58±1.21
Uniform	33.23±8.98	7.23±0.19	1.13±0.96	2.00±1.34

[a] n = 1122

A comparison of 'Remaining Values' to 'Imputed Values' shows that the normal distribution method was best at replicating the distribution of the remaining values and that the uniform method produced the most different mean.

The Mean (k-NN) method provided imputed values which were very similar in mean but slightly smaller in terms of standard deviation, which was expected as the missing value was replaced with the mean value, thereby lessening the effect of relative outliers at both ends of the distribution.

A comparison of 'True Values' and the 'Imputed Dataset' shows that the Mean (k-NN) method was best at re-creating the original values in the case of the Lserum variable, and performed equally or better than the other methods at replicating the mean for the three other variables. The normal distribution method more accurately matched the standard deviation of the true values in all cases except the Lserum variable. The Mean method was also successful at reproducing the mean of the 'True Values', however as the percentage of imputed values increased the reduction in standard deviation value became more pronounced.

Replacing values from a uniform distribution had the opposite effect of increasing the standard deviation in all cases even though the span was limited to 95% of the values. This method had the greatest influence on the mean of the values and resulted in the greatest change in variables with a higher percentage of replaced values. For three of the four variables considered, the final means were quite different from the original values. In the case of the normal distribution method, the imputed dataset statistics shifted towards those of the remaining values as the percentage of removed values increased. The uniform distribution method produced an imputed dataset which was the most different from the 'True Values' for all variables. Two mean values were larger and two were smaller, while the standard deviations were consistently larger.

The average percent error was calculated by comparing each original value to its imputed value and averaging the error over the number of values imputed using equations 5.3 and 5.4. The average error is acceptable for the Lserum variable with all imputation methods, but this is not the case for the other variables. The Normal Distribution and Uniform Distribution method present very large average error percentages, especially in the case of the Lurine variable. One possible cause is that most of the Lurine variables

are limited to a very small range, but the distribution contains some very significant outliers. These values, especially in the case of the Uniform Distribution method, could be responsible for much of the increase in average error compared to other methods. Table 6-3 provides additional information about the imputation methods. Pair-wise comparisons using the paired two tailed *t*-test were performed between the true and imputed values.

Table 6-3. Average percent error between known and imputed values

	Lbloodp	Lserum	Lurine	Po2fio2r
Mean (k-NN)	15.84 (%)	0.59 (%)	69.52 (%)	49.15 (%)
Mean	22.32 (%)	1.24 (%)	278.89 (%)	143.53 (%)
Normal Dist.	32.71 (%)	2.61 (%)*	635.57 (%)*	187.24 (%)
Uniform Dist.	37.16 (%)*	3.86 (%)*	615.14 (%)*	242.60 (%)*

* Significantly different from true values

The *t*-test showed no significant difference between the true and imputed values for the Mean (k-NN) and for the Mean approach ($p>0.05$). However, the mean percent error was consistently larger in the latter method. The results show that imputed values with the normal distribution were significantly different ($p<0.05$) in all cases except for the Lbloodp values ($p>0.05$). In the case of the uniform distribution method, all imputed values were significantly different ($p<0.05$) from the original known values. This method also had the greatest average percent error for each variable.

Although it was important that the replaced values had a group distribution similar to the original, the degree of similarity between each pair consisting of an artificially removed variable and its imputed value was also an important factor in the selection of a final imputation method. Table 6-3 shows that for the methods considered, the error was consistently smaller for the Lserum variable, likely because the variable had a very small standard deviation. For the three other variables considered, the error is considerably larger and had a relation to it's standard deviation and percentage of replaced values.

The Mean (k-NN) method used the relationship between the complete cases and the incomplete case under consideration to determine a replacement value. As a result of taking the mean value, the standard deviation of the variables decreased as values were

filled, however the imputed variables in the artificial dataset had a closer relationship to the original known values and the average percent error was lower.

In summary, for the incomplete SNAPPE-II variables, the Mean (k-NN) method had an averaging effect on the distribution of the variables, and was consistently the most accurate at imputing known values in the artificial dataset. Based on these results the Mean (k-NN) method was selected to impute unknown missing values in the three databases used in this research.

Imputing Missing Values

The SNAPPE-II variables with missing values along with the cumulative mortality rate were tracked in table 6-4 throughout the imputation process using the Mean (k-NN) method.

Table 6-4. Evolution of SNAPPE-II variables during imputation

	Lbloodp	Lserum	Lurine	PO2fiO2r	Mort (%)
Complete (n = 1122)	33.30±8.86	7.23±0.13	0.80±0.88	1.56±1.19	12.03
1 MV [a] (n = 2399)	33.79±8.63	7.29±0.13	0.84±0.83	1.61±1.14	10.30
2 MV [a] (n = 3567)	33.49±8.73	7.29±0.12	0.83±0.78	1.70±1.20	9.95
3 MV [a] (n = 5459)	34.19±8.83	7.30±0.12	0.83±0.74	1.82±1.19	8.48
4 MV [a] (n = 7095)	34.84±9.13	7.31±0.12	0.85±0.75	1.90±1.16	7.43
5 MV [a] (n = 8438)	35.26±9.36	7.31±0.11	0.86±0.74	1.95±1.13	6.87
6 MV [a] (n = 10492)	35.90±9.36	7.31±0.11	0.86±0.70	2.00±1.06	5.93
7 MV [a] (n = 11729)	36.22±9.43	7.32±0.11	0.87±0.70	2.02±1.03	5.47
8 MV [a] (n = 15384)	36.74±9.44	7.32±0.10	0.87±0.68	2.05±0.98	4.87
All known values	36.71±9.54	7.31±0.11	0.89±0.82	1.90±1.27	3.66

[a] MV indicates the maximum of missing values imputed per case

As expected because of the observations of table 6-4, the standard deviations of the variables in the imputed database were smaller than the standard deviations of 'All known values'. The last two rows of table 6-4 show the distribution of the variables with the maximum of eight imputed values ('8 MV') and without imputed values ('All known values'). As the number of missing values per case increased, the cumulative mortality percentage decreased as expected in a database increasingly comprised of survivors and less severely ill babies. This may account for the mean of certain variables shifting slightly towards a more normal value, as defined in the SNAP guidelines (appendix A).

The EPIC and CHEO databases were also imputed with the Mean (k-NN) method. The mean and standard deviation of each variable is listed in table 6-5. All babies in the EPIC database were born at under 32 weeks gestation while the CNN has cases with a gestational age up to 40 weeks. The patients in the CHEO database had a gestational age ranging from 24 to 41 weeks, with an average of 32.2. The table of outcome distributions (table 5-6) showed great variations in outcome and complication rates, therefore the variables were expected to be different (i.e.: lower birth weights, lower Apgar values, etc.) to reflect a sicker population.

Table 6-5. Mean and standard deviation for databases with imputed missing values

	CNN		EPIC		CHEO	
Variable	Mean	St.Dev	Mean	St.Dev	Mean	St.Dev
Lurine	0.87	0.65	1.35	0.74	2.52	1.98
Hrespr	58.58	29.07	52.24	13.84	62.23	17.57
Lplt	212.24	97.39	249.59	79.52	192.87	50.72
SGA	0.04	0.20	0.07	0.25	0.13	0.34
Hgluc	6.19	3.35	7.36	1.57	6.74	1.37
Lhema	48.08	8.56	41.80	4.91	42.65	3.54
Apgar5	7.83	1.73	7.44	1.63	7.31	1.86
Lbloodp	36.74	9.44	39.07	13.92	33.16	7.52
Hhema	49.31	8.68	42.56	5.49	44.79	3.12
Lgluc	3.43	1.58	3.97	0.78	5.09	1.91
birth weight	2339.48	1042.57	1192.90	466.15	1677.92	960.71
Hsodium	137.28	5.51	137.01	1.89	135.55	4.91
hitnrat	0.16	0.14	0.15	0.06	0.20	0.07
LWBC	14.78	8.91	12.97	3.32	11.40	3.47
LANC	8395.58	7724.69	7137.47	3085.17	6399.74	2963.18
Hbloodp	48.52	11.33	52.78	12.35	44.86	8.19
Lheartr	126.73	15.52	142.56	15.29	128.87	4.54
Hheartr	161.84	17.99	170.31	18.17	165.53	5.44
OI	6.65	10.74	7.08	6.60	7.87	7.97
Apnea	1.17	0.56	1.27	0.15	1.28	0.26
Ltempf	96.44	2.57	97.51	0.98	96.85	1.77
Seizure	1.05	0.29	1.15	0.61	1.10	0.44
Lsodium	135.17	3.71	135.15	1.55	134.68	2.15
Lpo2po2	42.88	31.10	55.33	28.16	61.45	15.39
Lserum	7.32	0.10	7.31	0.06	7.22	0.12
Po2fio2r	2.05	0.98	2.67	1.31	1.61	1.05
Hpco2	46.55	12.86	46.98	5.44	52.75	16.13

Again, the NTISS tables indicate that many babies in the EPIC database were incubated in transport, which may account for the EPIC database having the highest Ltempf mean. In future models, indicating whether or not a baby received help maintaining its body temperature could help differentiate between two babies who have the same value of Ltempf, though one had adequate temperature control and the other was helped with a warming chamber.

6.2 Determining the Reduced Variable Sets

Using the imputed CNN database, MSE ANN models were developed and their performance was tracked during series of variable eliminations. This section presents the performance of intermediate models and the importance of variables leading to the minimal variable sets. Removing variables, or parameter selection, is a common process in the optimization of ANN applications. Table 6-6 indicates how many cases the training, testing and validation sets contained, and table B-1 in appendix B lists the starting values for the control parameters. The classification performance of the MSE ANN models is presented one outcome at a time, beginning with mortality.

Table 6-6. Number of cases in each set

Outcome	Training	Testing	Validation	Prevalence of rare outcome
Mort	7186*	3021	4520	4.87 (%)
LOS ≥28	6043	3021	4520	27.31 (%)
DOV ≥8	6482*	3021	4520	14.20 (%)
BPD	6934*	3018	4520	8.12 (%)
IVH	7265*	3018	4520	3.72 (%)

* indicates the rare outcome was resampled to comprise 20% of cases

6.2.1 Mortality

On the next page, table 6-7 shows the performance of the non-linear MSE ANN when modeling the outcome of mortality, starting with the 27 input variables. The first two rows give the number of hidden nodes in the best performing structure and the epoch at which the listed performance was achieved. Five performance parameters are then provided and their values for the test set and for the average over ten validation sets is given. In the mortality classification experiments, the specificity of the test set was

maintained at 91.7±1.11% and at 92.39±0.75% for the validation set average. It was important to maintain a consistent specificity in both the testing and verification sets when comparing the models as sensitivity tends to increase whenever specificity drops as is seen on ROC figures.

Table 6-7. MSE ANN performance for mortality prediction during variable elimination

		27 variables	**24 variables**	**20 variables**	**18 variables**	**15 variables**
Best test epoch		357	116	524	1512	1097
Best structure		16	6	5	5	5
Log sensitivity	Best test	0.3824	0.4076	0.4903	0.4483	0.5097
	Validation average	0.3061	0.3128	0.3793	0.4584	0.4022
Sensitivity (%)	Best test	68.60	67.15	67.02	71.68	66.74
	Validation average	67.61	67.61	65.34	70.45	67.05
Specificity (%)	Best test	91.16	91.72	91.13	92.24	91.44
	Validation average	92.06	92.76	92.01	92.59	92.38
CCR (%)	Best test	90.40	91.00	90.43	91.59	91.40
	Validation average	90.87	91.54	90.71	91.51	91.15
ROC	Best test	0.8775	0.8434	0.8820	0.8883	0.8931
	Validation average	0.8900	0.8474	0.8834	0.9073	0.9028

The area under the ROC curve was above 0.85 for all models except the 24 variable model, and the 18 variable model displayed the best discrimination ability. The sensitivity and CCR were also at their highest in the 18 variable model. The average sensitivity increased when variables were removed from 27 down to 18, and then decreased in the 15 variable model. The increase likely occurred because variables which were acting as noise were removed until the final reduction which removed pertinent variables. In the 18 variable model the three least significant variables had nearly equal weighing. They were removed, resulting in the 15 variable model which had inferior performance.

To verify that the minimal dataset contained all 18 variables, the three lowest ranking variables were removed one at a time. When the Hheartr, LWBC and Seizure

parameters were removed, each model displayed problems maintaining the sensitivity achieved in the 18 variable model, confirming that all remaining 18 variables are critical.

Table 6-8 shows the importance (in decreasing order) of the variables during successive MSE ANN iterations. Italics denote the variables that were not retained for the following set of trials because of their low ranking.

Table 6-8. Relative importance of variables during variable reductions

27 variables		24 variables		20 variables		18 variables	
Apgar5	100	OI	100	Lplt	100	bthwht	100
Hrespr	77	Hrespr	60	LWBC	97	Ltempf	91
OI	67	Lurine	57	Hheartr	85	Apgar5	90
Lurine	65	SGA	52	apnea	85	OI	74
SGA	47	Apgar5	43	Apgar5	82	Lplt	74
bthwht	40	Hheartr	25	Ltempf	80	Hsodium	73
Lbloodp	36	Lserum	20	Lbloodp	74	apnea	72
Hhema	35	Ltempf	11	Po2fio2r	71	Hbloodp	72
Po2fio2r	30	Lplt	9	OI	70	Lhema	71
Lhema	26	seizure	8	Lurine	63	Lurine	61
Hgluc	24	Lhema	7	Hrespr	60	Hrespr	50
seizure	19	Lbloodp	7	Hsodium	60	Po2fio2r	46
Ltempf	16	Hsodium	6	seizure	55	Lserum	45
Hbloodp	15	LWBC	6	Lhema	45	SGA	31
Lserum	14	bthwht	6	Hbloodp	44	Lbloodp	29
LWBC	14	Hbloodp	6	Lserum	43	*Hheartr*	*20*
Lplt	12	Po2fio2r	6	bthwht	40	*LWBC*	*19*
Hheartr	11	apnea	5	SGA	33	*seizure*	*18*
lpo2po2	11	Hgluc	4	*Hgluc*	*33*		
Hsodium	10	Lheartr	4	*Lheartr*	*31*		
Lsodium	10	*Lsodium*	*3*				
apnea	9	*Lgluc*	*3*				
Lheartr	9	*Hhema*	*3*				
Lgluc	9	*lpo2po2*	*2*				
Hpco2	*6*						
hitnrat	*4*						
LANC	*2*						

Certain variables (OI, Apgar 5, Po2fio2r) remained quite high in all models, whereas others fluctuated. In the iteration with 20 variables, the last two of three variables were removed because SGA was a known important factor for mortality and had a weight equal to Hgluc. If the performance had decreased, a model with 19 variables (Hheartr removed only) would have been tried, but the performance of the model improved.

Table 6-9 compares the variables and weights of three day 1 mortality models: the Mean (k-NN) model developed in this thesis (also shown in Table 6-8), Ennett's Hybrid imputation model and the nine SNAPPE-II variables, shows interesting relationships. Both Hbloodp and Lbloodp are part of the minimal dataset obtained with the Mean (k-NN) imputation (table 6-8), though Hbloodp has more importance. This suggests that blood pressure regulation in the first 12 hours of life is an important indicator of health. Only one blood pressure variable is listed in Ennett's hybrid model [2003]. In fact Lbloodp and Seizure were the only two SNAPPE-II variables eliminated.

While birth weight ranked low (21) in Ennett's model, it was highest in both the Mean (k-NN) and SNAPPE-II models. This result makes clinical sense, knowing that premature (and thus low birth weight) babies account for 75%-85% of all perinatal deaths in Canada [McLaughlin et al. 1999].

Table 6-9. Comparison of variables and weights for day 1 mortality models

Mean (k-NN) (Table 6-8)		Ennett's Model [2003] (n=5102)		SNAPPE-II Variables (n=5102) [Ennett 2003]	
bthwght	100	po2fio2r	100	bthwght	100
Ltempf	91	Lurine	72	apgar5	62
apgar5	90	Lserum	36	seizure	43
OI	74	apgar5	33	Ltempf	41
Lplt	74	Lplt	29	Lserum	25
Hsodium	73	SGA	28	po2fio2r	20
apnea	72	Hsodium	28	SGA	15
Hbloodp	72	Hrespr	24	Lurine	3
Lhema	71	Hpco2	22	Lbloodp	1
Lurine	61	bthwght	21		
Hrespr	50	Lgluc	20		
Po2fio2r	46	Ltempf	18		
Lserum	45	Hbloodp	18		
SGA	31				
Lbloodp	29				
Hheartr	20				
LWBC	19				
seizure	18				

Another interesting observation is that the distribution of weights for the Mean (k-NN) model is more diffuse than Ennett's. Ennett's model had 2 variables with an importance of over 40, compared to 13 for the Mean (k-NN) model. The Po2fio2r and

Lurine had missing value percentages of 53 and 61; therefore it is possible that Ennett's weighted CBR imputation method had a greater effect on those variables.

The inclusion of the seizure variable in the Mean (k-NN) imputation model is interesting because it is a rare occurrence in the database (only 2.45% cases reported seizure events), however they were 4 times more likely to occur in mortality cases.

Objective 2 of this research was stated as follows in section 1.3:

> Compare the effect of weighted versus uniform imputation on the classification performance and relative variable importance in mortality models.

This new minimal dataset is larger than Ennett's (18 variables compared to 13). The disadvantage is that more variables are required to obtain a similar performance, however the same database may be used to predict other outcomes without the need to re-determine replacement values.

The classification performance of the same three mortality models is compared to a model with all SNAP variables in table 6-10 to address objective 2. The first column reports the final 18 variable mortality model obtained with Mean (k-NN) imputation showing a specificity above 90% and a ROC of 0.9073. The values for sensitivity and specificity in the second and third results columns are two other points on the 18 variable models' ROC curve. They represent models with the same discrimination ability. The two selected points are:

(1) 98.7 % specificity for a comparison to the SNAP and SNAPPE-II models

(2) 39.5 % sensitivity for a comparison to Ennetts' model

Table 6-10. Comparison of classification results for mortality models

	Mean (k-NN) [a]			Ennett	SNAP [b]	SNAPPE-
	Final	Point (1)	Point (2)	[2003] [b]		II [b]
Sensitivity (%)	70.45	29.3	**39.5**	**38.3**	23.0	26.6
Specificity (%)	92.59	**98.7**	97.6	98.9	**98.7**	**98.0**
CCR (%)	91.51			96.8	95.9	91.3
CP (%)	95.13			90.5	90.5	90.0
ROC	0.9073	0.9073	0.9073	0.8699		

[a] (n = 13584), [b] (n = 5102), [Ennett 2003]

The first performance point shows that when the specificity is very high, the Mean (k-NN) model performs better than the SNAP and the SNAPPE-II models in terms of sensitivity. Compared to Ennetts' model, the Mean (k-NN) model has lower sensitivity by 9%. This could be due to the fact that the CP is 5% higher, and that the database used by Ennett was much smaller.

The second performance point shows that when the Mean (k-NN) models' sensitivity is adjusted to Ennetts, the Mean (k-NN)'s specificity is only 1.3% lower.

In conclusion, the Mean (k-NN) imputation lead to a mortality model which determined a minimal variable set having a performance superior to models developed with the SNAP and SNAPPE-II variables. This outcome-independent database was therefore used to model other outcomes.

6.2.2 Length of Stay

The methodology used for the mortality outcome was applied to the length of stay to determine which variables acquired within the first 12 hours of admission are most relevant. Hospitalizations lasting 28 days or longer represent 27% of cases in the CNN, and are the most prevalent outcome considered in this thesis. The performance of the non-linear MSE ANN models is shown in table 6-11.

Table 6-11. MSE ANN performance for LOS prediction during variable elimination

		27 variables	**17 variables**	**16 (LWBC removed)**	**16 (Lserum removed)**
Best test epoch		1221	1770	1398	1648
Structure		8	9	29	7
Log sensitivity	Best test	0.2994	0.3553	0.3233	0.3580
	Validation average	0.2991	0.3348	0.3385	0.3547
Sensitivity (%)	Best test	71.88	72.00	70.67	69.92
	Validation average	69.33	70.46	68.01	69.11
Specificity (%)	Best test	90.03	90.12	90.00	90.07
	Validation average	89.37	89.72	89.59	89.67
CCR (%)	Best test	85.0712	85.17	84.74	85.20
	Validation average	84.51	84.48	83.74	84.81
ROC	Best test	0.8659	0.8811	0.8726	0.8776

Validation average	0.8595	0.8730	0.8695	0.8734

The first variable elimination resulted in a very similar though slightly better model in terms of sensitivity, CCR and ROC. This was expected as the variables removed were weighted with very small weights (<1). Removing either of the two weakest variables (LWBC and Lserum) in the reduced model decreased the performance. Birth weight was the most important variable in the model, and five of six parameters directly related to lung function were retained.

Table 6-12. Relative importance of variables during variable reductions

27 variables		**17 variables**	
bthwht	100	***bthwht***	100
Lbloodp	17	***Apgar5***	33
Hsodium	15	**Lbloodp**	32
seizure	12	Lhema	32
Hrespr	10	Hrespr	32
Hhema	6	OI	30
Lurine	6	*Hsodium*	28
apnea	6	Hhema	24
Lgluc	4	***Po2fio2r***	23
Lhema	3	***Lurine***	20
Apgar5	3	apnea	17
Po2fio2r	2	Hgluc	17
Hpco2	2	seizure	16
OI	1	*Lgluc*	14
Hgluc	1	*Hpco2*	11
Lserum	1	***Lserum***	8
LWBC	1	LWBC	6
SGA	<1		
Ltempf	<1		
Lheartr	<1		
Hheartr	<1		
LANC	<1		
hitnrat	<1		
Hbloodp	<1		
Lpo2po2	<1		
Lplt	<1		
Lsodium	<1		

Zernikow et al. [1999] used 40 parameters to classify lengths of stay which varied between 0 and 140 days into groups of ten days. Their ANN models obtained a ROC of 0.87 to 0.92 (no sensitivity or specificity provided), which was superior to multiple regression models. The shortest stays were consistently predicted more accurately then lengthier terms, and their case selection was limited to babies staying 140 days or less. The five most important factors outlined were gestational age, respiratory problems, birth weight, infections and metabolic problems [Zernikow et al. 1999]. Perhaps the inclusion of some of these variables in future MIRG MSE ANN models could improve the classification over what is provided by the minimal variable set.

6.2.3 Duration of Ventilation

Ventilators are expensive and limited in numbers therefore early identification of patients who will require more than 7 days of mechanical ventilation is important for resource and clinical management. The average ventilation duration was 4.70 days, with 49.87 % of babies requiring no artificial ventilation during their hospitalization. Correct prediction of an extended duration of ventilation presents some of the same problems as the extended length of stay because both rare outcomes represent events that occur multiple days after the variables were collected.

The specificity was kept stable at 90.91±0.60% in the test sets and 90.82±0.39% in the verification average. The performance of the MSE ANN for this outcome is presented in table 6-13.

Table 6-13. MSE ANN performance for classifying the duration of ventilation

		27 variables	**20 variables**	**16 variables**	**14 variables**	**13 (Lurine Removed)**
Best test epoch		960	1052	1196	830	1844
Structure		2	7	6	4	8
Log sensitivity	Best test	0.4771	0.3144	0.3856	0.3826	0.3912
	Validation average	0.4268	0.3235	0.3678	0.3675	0.3512
Sensitivity (%)	Best test	72.32	74.25	73.55	74.01	72.48
	Validation average	68.75	68.28	72.07	72.34	71.09

Specificity (%)	Best test	91.02	91.42	91.50	90.31	90.62
	Validation average	90.77	91.21	91.04	90.46	90.43
Classification CCR (%)	Best test	88.38	88.65	88.94	87.96	88.32
	Validation average	87.68	87.96	88.35	87.90	87.69
ROC	Best test	0.8570	0.8745	0.8970	0.8891	0.8755
	Validation average	0.8385	0.8652	0.8869	0.8816	0.8660

The sensitivity gradually increased as the number of variables in the models decreased from 27 to 14 variables and peaked at 72.34% in the verification set average. The ROC of all models was very good throughout the variable eliminations. The discrimination ability of the 14 variable model was 0.8816, which is only slightly inferior to the 16 variable model which obtained the highest value.

Table 6-14. Relative importance of variables during variable elimination

27 variables		**20 variables**		**16 variables**		**14 variables**	
bthwht	100	bthwht	100	bthwht	100	bthwht	100
OI	53	Hrespr	62	OI	61	OI	70
Lbloodp	32	Po2fio2r	41	apnea	47	Lbloodp	39
Hrespr	27	OI	35	Hrespr	42	seizure	39
LANC	21	Lurine	27	Hhema	41	Hheartr	29
apnea	17	Hheartr	27	seizure	33	Lgluc	28
Apgar5	16	hitnrat	25	Lgluc	33	Hrespr	26
Lserum	15	Lserum	25	Lbloodp	33	Hhema	19
lpo2po2	14	Lbloodp	22	Hheartr	32	apnea	19
Lplt	14	Hhema	22	Lurine	26	Po2fio2r	17
Hhema	12	apnea	22	Ltempf	26	hitnrat	15
LWBC	11	seizure	19	Apgar5	23	Lurine	11
Hheartr	11	Lgluc	19	Po2fio2r	21	Apgar5	11
Lurine	8	LANC	16	hitnrat	19	Ltempf	11
seizure	7	Apgar5	14	*LANC*	*12*		
Lgluc	7	Ltempf	13	*Lserum*	*8*		
SGA	6	*Lplt*	*9*				
Po2fio2r	5	*SGA*	*9*				
Ltempf	5	*LWBC*	*8*				
hitnrat	4	*lpo2po2*	*3*				
Lhema	*3*						
Hpco2	*3*						
Lheartr	*3*						
Hsodium	*3*						

Lsodium	*2*
Hgluc	*1*
Hbloodp	*<1*

Similarly to table 6-12 of the length of stay model, the distribution of the relative weights was diffuse during the first two rounds, which allowed many variables to be eliminated. This was not the case in the mortality model where many variables were emphasised equally and heavily. Only two variables were eliminated from the 16 variable model because the next variable weights were closely grouped together. When any of the three lowest ranked variables were removed from the set of 14 variables the sensitivity decreased. The model with Lurine removed performed best, but was still inferior to the 14 variable model.

Premature babies are likely to have immature lungs and difficulty breathing on their own. The importance of birth weight in all the models reflects this fact. Following birth weight the highest weighted variable is oxygenation index (OI) which provides information about the degree of mechanical ventilation. In future models, the ventilation status at 12 hours post admission would be an interesting variable to add. It could provide information on whether long term ventilations are more likely to start at admission or later during the hospitalization and help improve classification results.

Summary of outcome models: The best performing models had 18, 17 and 14 variables in the mortality, length of stay and ventilation outcomes respectively. The highest sensitivity was 74.01%, obtained in the ventilation model, which is only 2 percentage points above mortality and length of stay results.

In the three outcomes considered, there is no direct relationship between the number of variables in the minimal variable set and the average number of days elapsed between admission and the outcome; an outcome that occurs on day 28 does not necessarily require more variables than an outcome that occurs on day 8, to model. Eight variables were retained in all three models (Apgar5, apnea, birth weight, Hrespr, Lbloodp, Lurine, OI, seizure), and three variables (Lheartr, LANC and Lpo2po2) were eliminated from the three models.

The minimal variable set models were applied to the 59 cases of the EPIC database in order to assess their generalization ability and enable a comparison to previous mortality and length of stay models. Table 6-15 shows the classification performance as well as the number of cases in each category of the confusion matrix.

Table 6-15. MSE ANN models applied to EPIC

Outcome	Positives	TP	FP	Negatives	TN	FN	Sens. (%)	Spe. (%)	CCR (%)	CP (%)
Mortality	8	5	5	51	46	3	62.50	90.20	86.44	86.44
LOS	22	11	6	37	31	11	50.00	83.78	71.19	62.71
DOV	15	9	9	44	35	6	60.00	79.55	74.58	74.58

The ventilation outcomes modeled in previous MIRG theses were of ventilation durations above 24 hours, whereas this thesis uses 8 days as cut off, making comparisons with previous models impossible.

The first observation to make from table 6-15 is that the CCR is equal to or above the CP in all cases. For the mortality outcome the CNN model performed equally well on EPIC and had a greater sensitivity than was reported for the physicians during the first survey where the models had correctly classified two deaths and the physicians four.

For LOS and DOV, the lower specificity compared to the validation sets of the CNN was expected because the EPIC cases were "expected" to have long hospitalizations but perhaps faired better due to the protocols applied. The higher number of false positives in the DOV outcome may signal a similar situation, resulting from the improved protocol.

Importantly, these models applied to EPIC provided equal performance to both the committee of classifiers developed by Rybchynski [2005] in the case of mortality, and superior performance in sensitivity, specificity and CCR in the LOS model shown earlier in table 4-3. In comparison, Qi's mortality model in table 4-5 obtained a specificity of 98%, but a sensitivity of 25%, resulting in a CCR slightly below the CP. For the LOS outcome, all three performance indicators were 63%. These values were higher than the doctors' predictions however the models developed for this thesis show an increase of 20% in terms of specificity and 8% in CCR, though the sensitivity is lower by 13% compared to Qi [2005].

Since the models obtained with the Mean (k-NN) imputation achieved classification performances that were equal to or better than models obtained in previous attempts, the complications were also modeled, beginning with BPD.

6.2.4 Bronchopulmonary Dysplasia

The first complication selected was BPD, which had a prevalence of 8.12 % in the CNN. The MSE ANN's classification results are in table 6-16.

Table 6-16. MSE ANN performance for BPD classification during variable elimination

		27 variables	**24 variables**	**19 variables**	**18 variables**
Best test epoch		1765	624	603	769
Best structure		6	5	8	13
Log sensitivity	Best test	0.4553	0.4411	0.4538	0.4554
	Validation average	0.4621	0.4531	0.4951	0.4767
Sensitivity (%)	Best test	79.51	78.28	79.10	78.11
	Validation average	77.74	77.77	80.09	78.24
Specificity (%)	Best test	90.60	90.05	90.05	90.18
	Validation average	90.10	90.10	89.94	90.67
CCR (%)	Best test	89.70	90.32	89.17	89.29
	Validation average	89.16	89.12	89.16	89.68
ROC	Best test	0.8958	0.8706	0.8825	0.8811
	Validation average	0.9050	0.8767	0.8868	0.8785

In the four groups of variables tested for BPD, the specificity of the test and validation sets was kept within a tight range: 90.325±0.28% and 90.305±0.37% respectively. The area under the ROC curve was satisfactory for all models, with the 19 variable model discriminating better than the other reduced variable models.

Table 6-17. Relative importance of variables on BPD classification

27 variables		24 variables		19 variables	
bthwht	100	bthwht	100	Lgluc	100
LWBC	67	Hrespr	24	Hheartr	18
Po2fio2r	65	OI	4	Hrespr	13
Hrespr	64	Lhema	4	Lbloodp	7
Apgar5	57	Lbloodp	4	Apgar5	7
Lbloodp	53	Hsodium	4	Lserum	5
Lplt	50	Lheartr	4	Hgluc	5
Hhema	49	Hhema	4	bthwht	3
Lgluc	49	Po2fio2r	4	Hbloodp	3
Lhema	47	Hheartr	3	LWBC	3
Hbloodp	46	Lserum	3	LANC	3
Lheartr	45	Hbloodp	2	Hhema	2
OI	42	Lplt	2	Lhema	2
hitnrat	38	seizure	2	Lplt	2
Hgluc	36	Lgluc	1	Lheartr	2
Hsodium	35	LWBC	1	seizure	1
Hheartr	34	LANC	1	Hsodium	1
seizure	33	Apgar5	1	Po2fio2r	1
LANC	33	Hgluc	1	OI	<1
Lurine	26	*lpo2po2*	*<0.5*		
Hpco2	26	*Hpco2*	*<0.5*		
Lserum	25	*Ltempf*	*<0.5*		
lpo2po2	25	*Lurine*	*<0.5*		
Ltempf	23	*hitnrat*	*<0.5*		
Lsodium	*12*				
apnea	*9*				
SGA	*8*				

After the first elimination, five variables were weighted under 0.5 (half the size of the next set of five with a weight of 1) so they were removed from the 24 variable model. The OI variable was not heavily weighted in the 19 variable model, and lost rank position compared to the 27 variable model but was important for the correct classification of certain cases as the sensitivity of the model decreased when it was removed. Though the sensitivity of the validation set reached 80% in the final model, there was little improvement in the sensitivity of successive models. This may be a result of using a limited type of variables (physiologic) and using babies of all birth weights and gestational ages.

BPD is a condition that occurs more frequently in babies of lower gestational age, therefore it is possible that a dataset limited to babies with a gestational age of 32 weeks or less would result in improved classification [Lee 2006].

6.2.5 Intraventricular Hemorrhage

The second complication was the diagnosis of a severe IVH which occurred in 3.72% of cases. The classification results during variable reductions are shown in table 6-18. Specificity was maintained at 90.48±0.60% in the test set and 90.31±0.33% in the validation sets. The percentage of IVH diagnosis is so small that each of the ten validation sets only contained sixteen positive cases. The sensitivity increased as variables were removed until 23 remained, then quickly decreased. For these models, the ROC was consistently lower than for the other outcomes considered. Only the 23 variable model had a ROC above 0.85 for both the testing and validation sets. Though the sensitivity increased with variable eliminations, it remained lower than in models for other and more prevalent outcomes.

Table 6-18. MSE ANN performance for IVH classification during variable elimination

		27 variables	**24 variables**	**23 variables**	**22 variables**
Best test epoch		281	383	1366	911
Best structure		6	7	13	21
Log sensitivity	Best test	0.3278	0.4017	0.3384	0.1447
	Validation average	0.1799	0.2526	0.1936	0.1502
Sensitivity (%)	Best test	50.52	52.99	54.37	48.52
	Validation average	50.63	51.95	53.13	47.91
Specificity (%)	Best test	91.60	90.69	92.12	89.88
	Validation average	91.10	89.98	91.28	90.21
CCR (%)	Best test	83.39	83.94	91.09	81.61
	Validation average	89.99	88.90	89.93	88.72
Area under ROC	Best test	0.8465	0.8494	0.8610	0.8270
	Validation average	0.8521	0.8599	0.8538	0.8464

Chien et al. [2002] stated that severe IVH usually occurs within 72 hours of birth. The information contained in the CNN does not allow for a separation of babies which showed signs of severe IVH shortly after birth and those who took longer to develop symptoms. It is possible that models integrating day 3 information would be able to predict severe IVH with greater accuracy. IVH is usually diagnosed with a cranial ultrasound, and this is not typically performed within the first 24 hours [Bariciak 2006].

Table 6-19 shows the relative importance of SNAP variables in the classification of cases with a severe IVH diagnosis. The variables in this final model were more closely weighted than in the 27 and 24 variable models.

Table 6-19. Relative importance of variables during variable elimination

27 variables		**24 variables**		**23 variables**	
bthwht	*100*	***bthwht***	100	***bthwht***	100
Lplt	57	**Lbloodp**	87	Hrespr	100
Hrespr	53	*Lplt*	69	*Lplt*	85
Lbloodp	50	hitnrat	68	***Apgar5***	82
LWBC	48	Lhema	67	*Hbloodp*	79
Lgluc	*48*	LWBC	63	LWBC	79
Po2fio2r	45	OI	55	***SGA***	78
Lheartr	*28*	Lheartr	54	Lsodium	77
Hgluc	*26*	Hrespr	53	Lhema	76
Hpco2	25	***Po2fio2r***	50	Hgluc	75
hitnrat	22	***SGA***	49	LANC	73
Hheartr	*21*	***Apgar5***	48	***Po2fio2r***	72
Lhema	*20*	*Lgluc*	44	apnea	70
LANC	*19*	Lsodium	43	Lheartr	69
Lurine	*19*	*Hsodium*	42	*Lgluc*	66
Hbloodp	15	LANC	41	hitnrat	64
Hsodium	10	*Hbloodp*	41	**Lbloodp**	63
Lsodium	*9*	***Lurine***	38	Hhema	62
Hhema	*7*	Hhema	37	OI	60
SGA	6	***Lserum***	35	*Hsodium*	57
apnea	6	*Hpco2*	33	***Lserum***	53
Apgar5	*5*	Hgluc	33	***Lurine***	52
Lserum	*5*	apnea	31	*Hpco2*	*43*
OI	4	*Hheartr*	*29*		
seizure	*3*				
lpo2po2	*3*				
Ltempf	*0*				

Both Hbloodp and Lbloodp carried strong weights, which correlates with Annibale and Hills' [2006] finding that severe IVH may be caused by problems with blood pressure regulation. As well, the presence of Hrespr, apnea and three ventilator parameters (Po2fio2r, OI and Hpco2) draw a parallel with the observation by Chien et al [2002] regarding the asynchronicity between mechanical and spontaneous breathing. An IVH occurs almost exclusively in very and extremely low weight babies; a model limited to babies below a certain weight may reveal a smaller minimal variable set.

Summary of complication models: As variables were removed, an improvement in classification was observed because variables that were less relevant to the outcome or acted as noise were removed. The modified Garson-Goh method was able to classify the variables in order of relative importance. The third objective of this thesis:

Determine a minimal variable set for each outcome considered using a single database.

was reached. Results obtained were in general superior or equal (for DOV and LOS and complications) to those obtained with the hybrid imputation [Rybchynski 2005], [Qi 2005], confirming that the uniform weights approach can successfully model outcomes other than mortality. The minimal datasets for the complications contained more variables than those of the first three outcomes possibly because the SNAP variables were selected to identify illness severity rather than particular illnesses [Richardson et al. 2001]. Suggestions were made throughout this thesis' results section of variables to assess which may increase the classification accuracy. For example, it is likely that retaining only cases with a gestational age below 32 weeks would result in a model with higher sensitivity, though it would limit the application of the model to similar cases (not applicable to babies with a gestational age above 32 weeks).

The number of hidden nodes in the best performing structures did not consistently increase, stabilize or decrease as variables were removed, which emphasizes the need for experimenting with multiple structures. Figure 6.1. plots the improvement in sensitivity versus the number of remaining variables during the variable elimination for all

outcomes. Each data point represents the sensitivity of the best performing structure out of the 2n+1 structures attempted, where n is the number of variables remaining.

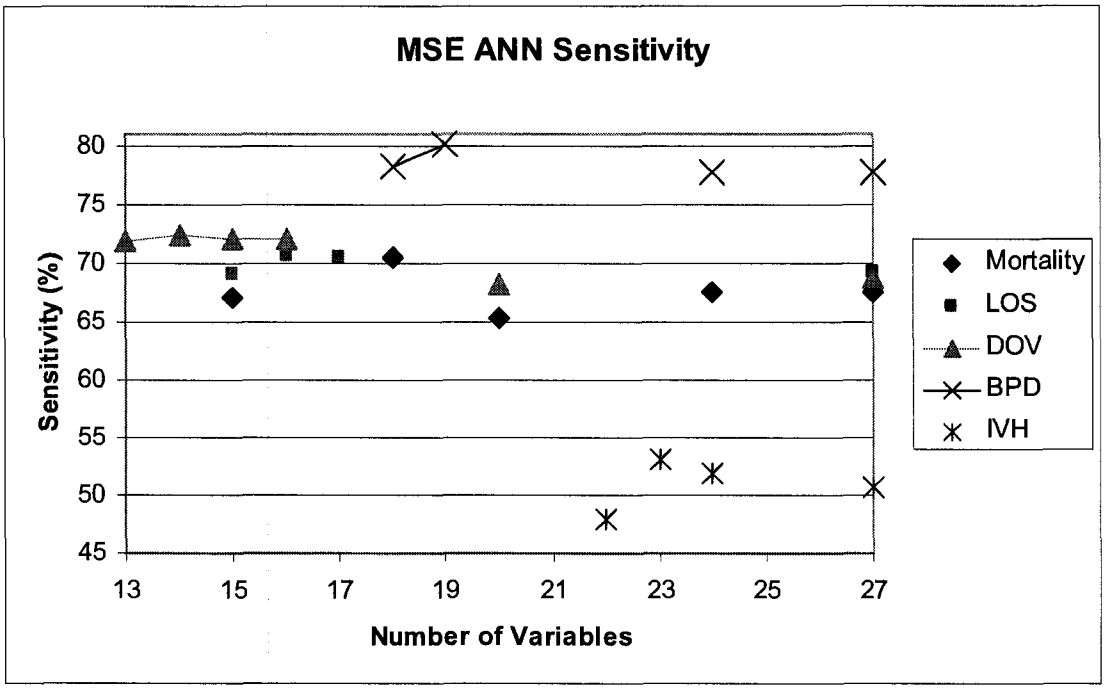

Figure 6.1. Effect of variable reductions on MSE ANN sensitivity

For all outcomes considered the non-linear models performed better than the linear models (structures with no hidden nodes), suggesting that the 27 SNAP variables and outcomes considered do have a non-linear relationship though some linear component is also discernable.

Zamboni [2007] used the CNN database in her study of the time-varying properties of the data on the discrimination ability of MSE ANN and ML ANN models as part of her profile of risk factors for NEC. Her analysis of day 1 data led to a 17 variable minimal dataset which classified cases at 82.61% sensitivity and 82.24% specificity by using a non-linear network with 30 hidden nodes. The MSE ANN was trained with 13133 cases as 415 cases were removed due to missing NEC information or because the baby's gender, which is often reported in medical literature at a factor for NEC, was unknown or listed as ambiguous.

In the outcomes considered, six or more SNAPPE-II variables were present in the minimal variable sets as summarized in table 6-20.

Table 6-20. Summary of minimal variable sets for all outcomes considered

SNAP Variables	Mortality	LOS	Vent	BPD	IVH	NEC
bthwght	√	√	√	√	√	√
Ltempf	√		√			√
Apgar5	√	√	√	√	√	√
OI	√	√	√	√	√	√
Lplt	√			√	√	
Hsodium	√	√		√	√	√
apnea	√	√	√		√	√
Hbloodp	√			√	√	
Lhema	√	√		√	√	
Lurine	√	√	√		√	√
Hrespr	√	√	√	√	√	√
Po2fio2r	√	√	√	√	√	√
Lserum	√	√		√	√	√
SGA	√				√	
Lbloodp	√	√	√	√	√	
Hheartr	√		√	√		
LWBC	√	√		√	√	√
seizure	√	√	√	√		√
Hhema		√	√	√	√	
Hgluc		√		√	√	√
Hpco2		√			√	√
Lgluc		√	√	√	√	√
hitnrat			√		√	√
Lheartr				√	√	
LANC				√	√	
Lsodium					√	
Lpo2po2						

The NEC model also includes a binary variable which indicates if the baby's gestational age is above or below 37 weeks to identify a baby as pre-term.

6.3 Risk Stratification

The fourth objective of this thesis was to:

> Use neural network learning with the maximum likelihood estimation to estimate the risk stratification for six important clinical outcomes and classify the risk into three categories: low (0% - 20%), moderate (21% - 74%) and high (75% - 100%).

The estimated risk stratification was created by using ML ANNs to create models of the estimated risk stratification. During the structure trial process, the maximum log-sensitivity index was again used as a stopping criterion because most of the outcomes are of low prevalence in the CNN database. The training and testing sets contained the same number of cases for each outcome because no resampling was performed.

Table 6-21. Complete, training and test set statistics for risk models

	Complete Set		Training Set		Testing Set		# of
	# cases in total	% rare outcome	# cases in total	% rare outcome	# cases in total	% rare outcome	structures (2n+1)
Mortality	13584	4.87	9056	4.87	4528	4.86	37
LOS	13584	27.31	9056	27.46	4528	27.01	35
DOV	13584	14.20	9056	14.33	4528	13.94	29
BPD	13584	8.12	9056	8.04	4528	8.28	39
IVH	13584	3.73	9056	3.77	4528	3.64	47
NEC	13133	2.38	8755	2.35	4378	2.43	35

The structures were rated on their p-value (>0.05) and ROC value (>0.85). The CCR was considered relevant for outcomes that are more frequent (LOS and DOV). The output of the network indicates the estimated degree of risk associated with the occurrence of the rare outcome. An output of 1 indicates that the outcome of mortality, for example, is expected with a high level of certainty while an output near 0 indicates a high level of certainty that mortality *will not* occur. The classification results are later presented for a cut-off value of 0.5 on the output node.

The optimal learning rate varied between the models and required many trials to pinpoint. In general, the structures took a larger number of epochs to converge than did

the MSE ANN; likely because the training set did not benefit from resampling which is known to decrease training time by repeatedly presenting the rare outcome cases. The pertinent details of the best performing structure for each outcome are given in table 6-22, and table B-2 in appendix B lists the starting values for the control parameters.

Table 6-22. Results for the best performing structure for each outcome (test set)

Outcome	Structure	Learning Rate	Best Epoch	Sensitivity (%)	Specificity (%)	CCR (%)	ROC Area
Mortality	15	5.014e-05	3160	28.25	99.21	95.72	0.8906
LOS	15	2.886e-04	4467	67.92	91.05	84.81	0.8801
DOV	26	8.834e-05	7000	51.51	96.23	89.99	0.9003
BPD	10	6.623e-04	1423	28.27	98.03	92.25	0.9081
IVH	0	0.0030418	1773	3.03	99.82	96.29	0.8188
NEC	9	0.0024590	1234	6.54	99.86	97.65	0.8173

The sensitivity values are lower than for the MSE ANN models; partly responsible is the much higher specificity. The specificity of the final models is highest in the outcomes with very unbalanced outcomes (mortality, IVH and NEC), whereas it is lowest for the more prevalent ones (LOS and DOV). This results from the calculation of the error function which will favour the most prevalent class, not the lack of resampling in the training set. The CCR of the estimated risk stratification showed some improvement over that of the MSE ANN models for all outcomes except the LOS model where the CCR is <1% lower. The ROC of each model is very comparable to its MSE ANN model.

The low prevalence of the rare outcome affects the ability of the algorithm to fit a model which has a sensitivity value comparable to the MSE ANN models when the cut off value of the output node is set to 0.5. Fortunately, this method, contrary to the MSE ANN used in 6.2, allows for a modifiable cut-off value of the output node to let the model favour either sensitivity or specificity. A models' sensitivity can be increased by lowering the cut off value since the estimated risk values for the rare outcomes are generally higher than those of the common outcomes. An attempt to steer the error calculation towards the under-represented class by using a weight factor resulted in models which were not acceptable ($p < 0.05$) [Zhou 2006].

Table 6-23 shows the results of the H-L goodness of fit test (introduced in section 3.1.2.3) for each dataset, the degree of freedom and the value of the log-likelihood in the stopping criterion for each model of table 6-22. It serves to demonstrate the models' goodness of fit on the data. When fewer than 5 cases are in any of the ten probability group, they are combined with those of a neighbouring group to avoid biasing the p-value and affecting the assessment of the model's goodness of fit [Zhou 2006]. Most often the lack of common outcome cases in the higher risk groups, due to a high specificity, is responsible for the reduction in the degree of freedom.

Table 6-23. Hosmer-Lemeshow test results for all datasets

	Set	C-Value	Deg. of Freedom	Log Likelihood	p-Value
Mortality	Train	6.992	8	-1034.81	0.10795
	Test	8.2329	8	-572.305	0.41106
	All	6.3205	8		0.61138
LOS	Train	8.2126	8	-3254	0.095023
	Test	12.2098	8	-1675.01	0.14209
	All	14.027	8		0.081063
DOV	Train	9.0773	8	-2179.65	0.08327
	Test	7.6089	8	-1108.39	0.47258
	All	7.985	8		0.43494
BPD	Train	6.7913	6	-1528.03	0.096622
	Test	11.0442	6	-792.68	0.087021
	All	11.9741	6		0.06255
IVH	Train	5.3765	4	-1102.62	0.0914
	Test	6.1062	3	-580.031	0.10655
	All	6.6591	5		0.24725
NEC	Train	6.263	4	-1093.77	0.078349
	Test	12.3384	4	-586.037	0.057004
	All	12.3708	5		0.060454

Since the first three outcomes observed were more prevalent, there were generally more cases in the higher estimated risk categories, leading to a model with a higher degree of freedom. The BPD model has six degrees of freedom even though the outcome is more prevalent than mortality, which has eight. This may be attributed to the established relationship between the SNAP variables and mortality.

Table 6-24 shows the distribution of observed deaths and observed survivors in each probability group of the mortality model whose properties were listed in tables 6-22 and 6-23. The number of observed deaths in each probability group is obtained by sorting the estimated risk of the test cases into probabilities groups, and the number of expected deaths is determined by the model. The parity between the numbers in the 'Expected' and 'Observed' columns speaks to the goodness of fit of the model. We can see that in general, there are more cases in the groups of lower probability. This is because there are dramatically more survivors (and thus a lowered likelihood of mortality) and due to actual mortalities classified at lower likelihoods nearly two thirds of the time because the sensitivity of the test set is 30%.

Table 6-24. The estimated risk stratification of the mortality outcome (test set)

Probability Group	Expected mortalities	Observed mortalities	Expected survivors	Observed survivors	Total Cases in Group
<0.1	61.7	74	3960.3	3948	4022
0.1-0.2	31.8	30	188.2	190	220
0.2-0.3	23.4	26	72.6	70	96
0.3-0.4	18.1	16	33.9	36	52
0.4-0.5	18.3	14	22.7	27	41
0.5-0.6	17.3	17	14.7	15	32
0.6-0.7	13.8	14	7.2	7	21
0.7-0.8	21	19	7	9	28
0.8-0.9	8.5	7	1.5	3	10
0.9-1.0	5.7	6	0.3	0	6

The majority of cases (89%) were classified in the <0.1 probability group. Two reasons explain this: the sensitivity of the model is low, which means fewer "observed mortalities" are expected to be in higher risk brackets, and the survival outcome is much more prevalent and was classified correctly in over 99% of cases. The number of observed mortality cases increases from the 0.5-0.6 risk group upwards. At that point there are more mortalities than survivor cases in all probability groups which is significant considering the discrepancy in prevalence of these outcomes. It shows that very few survivors will be classified in the higher risk brackets. The estimated risk stratification of the other outcomes is in appendix C.

In conclusion, it appears that first day of life data contains sufficient information to estimate the degree of estimated risk of important clinical outcomes and complications with very high specificity and reasonable sensitivity in most cases. The results are potentially useful for doctors, parents and hospital administrators to plan resources, and for parents to cope with the stress and uncertainty of having a sick baby.

Risk Groups

To meet the second criteria of the fourth objective, the cases of the CNN were sorted into three categories according to the estimated risk stratification. Table 6-25 shows the number of cases in the high, moderate and low risk categories for all outcomes. In the second column, the outcome '1' represents the rare outcome and '0' the common outcome, followed by the number of cases and their percentage in the CNN database.
In the Risk Category columns, the number preceded by '*n* =' represents the number of cases and the following percentage represents the fraction of '1' or '0' cases which were classified in that category (i.e. 282 mortalities were classified in the high estimated risk category, representing 43% of all mortalities).

Table 6-25. Positive and negative cases of the CNN according to estimated risk category

		Risk Category		
		High	Moderate	Low
Mortality	1 (*n* =661, 4.87%)	***n* =282, 43%**	*n* =299, 45%	*n*=80, 12%
	0 (*n* =12923, 95.13%)	*n* =12, 0.09%	*n* =456, 4%	*n*=12455, **96%**
LOS	1 (*n* =3710, 27.31%)	*n* =1370,37 %	*n*=1808, 49%	*n* =532, 14%
	0 (*n* =9874, 72.69%)	*n* =277, 2%	*n*=1939, 20%	*n*=7658, 78%
DOV	1 (*n* =1929, 14.20%)	*n* =427, 22%	*n*=1439, 75%	*n* =63, 3%
	0 (*n* =11655, 85.80%)	*n* =91, 1%	*n*=5977, 51%	*n*=5587, 48%
BPD	1 (*n* =1103, 8.12%)	*n* =17, 2%	*n* =821, 74%	*n* =265, 24%
	0 (*n* =12481, 91.88%)	*n* =17, 0.14%	*n* =1137, 9%	*n*=11327, 91%
IVH	1 (*n* =506, 3.73%)	*n* =3, 1%	*n* =135, 27%	*n* =365, 72%
	0 (*n* =13078, 96.27%)	*n* =0, 0%	*n* =371, 3%	*n*=12707, 97%
NEC	1 (*n* =312, 2.38%)	*n* =4, 1%	*n* = 89 , 29%	*n* = 219, 70%
	0 (*n* =12821, 97.62%)	*n* = 0, 0%	*n* = 333, 3%	*n* =12488, 97%

The high risk category represents cases which were attributed an estimated risk percentage of 75 or more, the moderate category represents 21-74% and a low estimated risk is 20% or less. The rightmost column of table 6-25 shows that the common outcome is primarily classified as low risk (96% of survivors for example) due to the specificity being quite high. The rare outcome cases are distributed in the other categories with different concentrations depending on the sensitivity achieved by the model. Nearly equal numbers of mortalities were classified as high and moderate risk while one tenth of cases were classified as low risk. In the ventilation model however, the majority of positive cases were classified in the moderate category.

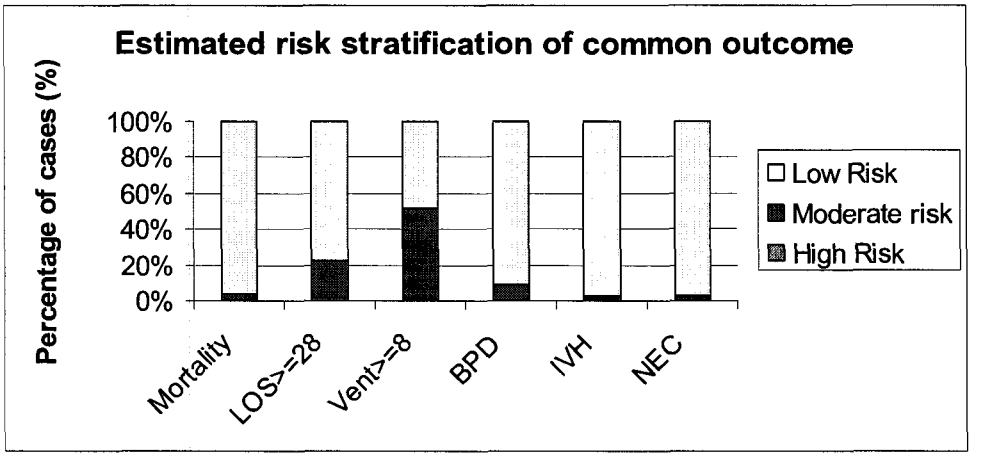

Figure 6.2. Common outcomes according to estimated risk category

Figure 6.2 is a graphical representation of the distribution of the common outcomes using the data of table 6-25. Very few cases were classified in the high risk category, which results of maintaining a high specificity. Only the ventilation model had more common outcomes (short term ventilation) in the moderate risk category than in the low risk category.

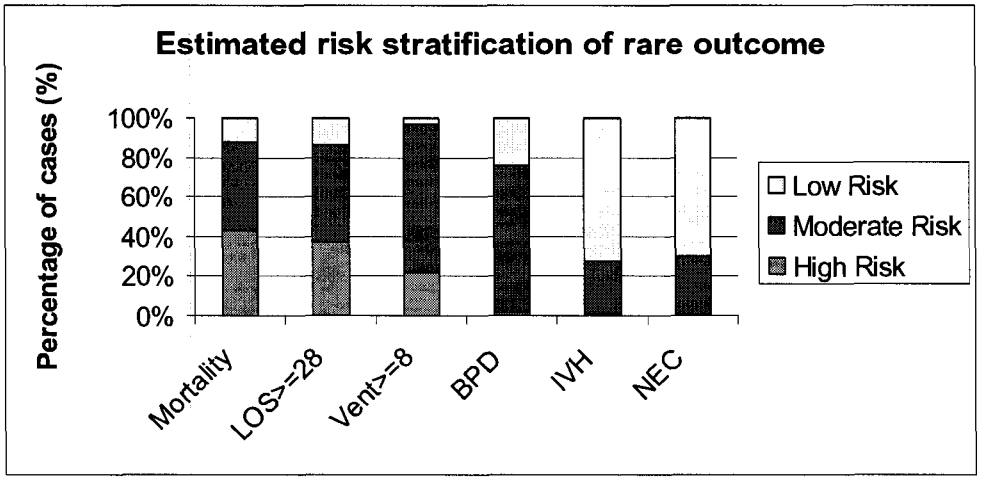

Figure 6.3. Rare outcomes according to estimated risk category

As seen in Figure 6.3, the rare outcomes were most often classified as moderate risk, perhaps because it has largest range and because the low sensitivity of the model prevents many cases from having higher estimated risks. The mortality model has the highest percentage of positive cases in the high risk category (43%), followed closely by LOS at 37%. The ventilation outcome has the highest percentage of cases, 75%, classified as moderate risk and the IVH at 27% has the lowest. The IVH model, which had the lowest sensitivity, has the highest percentage of positive cases, 72%, in the low risk category.

Having a narrow range for the low and high risk categories and one over twice as wide for the moderate category, as chosen by the neonatologists, results in a lower percentage of rare outcomes in the high risk category than if the categories had been divided evenly over the 0-1 range.

Extending Models to CHEO Database

The estimated risk stratification models were applied to the CHEO database in preparation for the clinical survey. The results are presented in table 6-26.

Table 6-26. The cases of CHEO (n = 60) divided into three estimated risk categories

		Risk Category		
		High	Moderate	Low
Mortality	1 (n =11) (18.33%)	n =2, 18%	n =5, 45%	n =4, 36%
	0 (n=49) (81.67%)	n =1, 2%	n =5, 10%	n =43, 88%
LOS	1 (n =21) (35%)	n =8, 38%	n =11, 52%	n =2, 10%
	0 (n=39) (65%)	n =12, 31%	n =10, 26%	n =17, 43%
DOV	1 (n =27) (45%)	n =5, 19%	n =13, 48%	n =9, 33%
	0 (n=33) (55%)	n =1, 3%	n =10, 30%	n =22, 67%
BPD	1 (n =2) (3.33%)	n =0, 0%	n =2, 100%	n =0, 0%
	0 (n=58) (96.67%)	n =0, 0%	n =21, 36%	n =37, 64%
IVH	1 (n =3) (5%)	n =0, 0%	n =3, 100%	n =0, 0%
	0 (n=57) (95%)	n =0, 0%	n =3, 5%	n =54, 95%
NEC	1 (n =7) (11.67%)	n =0, 0%	n =4, 57%	n =3, 43%
	0 (n=53) (83.33%)	n =0, 0%	n =3, 6%	n =50, 94%

Again, the majority of mortalities were classified in the moderate risk group and the vast majority of survivors were classified as low risk. In the case of the LOS outcome, the rare outcome was classified in a manner similar (slightly improved) to the CNN, while the common outcome was more often classified in the high risk category. When considering a limited number of cases, occurrences of misclassification may not be representative of the model's performance on a larger dataset.

The short term ventilations (n = 27) were very well classified considering this population had 58 out of 60 cases on a ventilator at some point during their stay, compared to under 50% in the CNN.

There were no BPD cases in the low risk category, though some degradation in cases without the diagnosis of BPD (n = 58) with 36% classified as moderate. This is likely a reflection of the case selection criteria which greatly favoured cases where babies

showed signs of respiratory distress. BPD was diagnosed if a baby required supplemental oxygen at 36 weeks corrected age [CNN 2004], which may be an argument for the addition of the **corrected** gestational age as an input variable. The need for supplemental oxygen is tracked in the NTISS table.

The NEC model was the only complication which used the gestational age as an input instead of the small for gestational age status, and the classification may have benefited from this variable substitution.

Because the inputs of the models were limited to the SNAP variables, the three complications considered were not as satisfactorily classified as the other outcomes. This would suggest that the SNAP variables may not be sufficiently indicative of these particular complications. Now that minimal variable sets have been established, the link between the individual complications and other scoring systems such as NTISS could be analyzed. A possible option would be to incorporate the gestational age of the baby at the end of the predicted ventilation duration into the BPD classification.

6.4 Integration of Components

In this research, models were developed with the CNN database and subsequently applied to a special set of cases collected at CHEO. Many steps were required to integrate the components detailed in sections 6.1-6.3 with the interface and accomplish the fifth objective of this thesis. The three primary steps were: preparing the data, conducting the survey, and analysing the results.

The stages of the preparation steps are numbered Figure 6.4. First, the selected variables were abstracted from archived charts at CHEO (step (1) in Figure 6.4) and entered into an Excel spreadsheet. Variables were then classified as belonging to the mother, baby or the SNAP list (2). Those in the latter category were modified to match the units in the SNAP guidelines as described in section 5.1.3. Using the caselink field that identified each case individually, the six outcomes considered were determined to have occurred or not by reading the admission and discharge summaries.

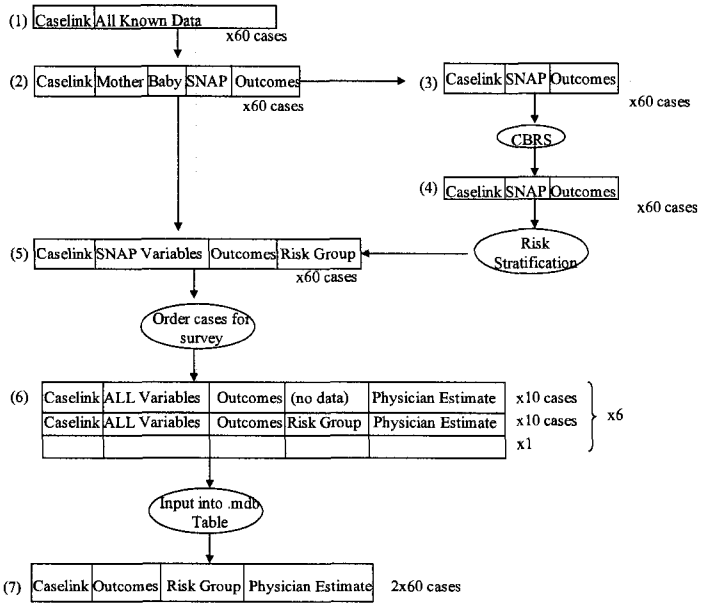

Figure 6.4. Diagram of component integration in preparation for the survey

From (2), the caselink and SNAP variables were processed with the CBR system to impute missing values (3). The risk stratification estimates developed in section 6.3 were then applied to the complete CHEO cases (4), and each outcome was labelled with the appropriate estimated risk category (5).

Because the projected number of participants was six, it was decided to present each participant with twenty cases. This enabled each case to be presented twice over the course of the survey: once to a physician without the estimated risk category provided, and once to a second physician this time with the estimated risk category displayed. Accordingly, the database was organized to present one group of ten cases without risk estimates, followed by ten new cases with estimates, followed by one blank case (all fields empty) to signal the end of the survey to participants (6). An inspection of the database verified that rare outcomes were divided as uniformly as possible and that the 120 cases were arranged as planned. This database was copied into an MS Access table that populated the fields in the interface presented below.

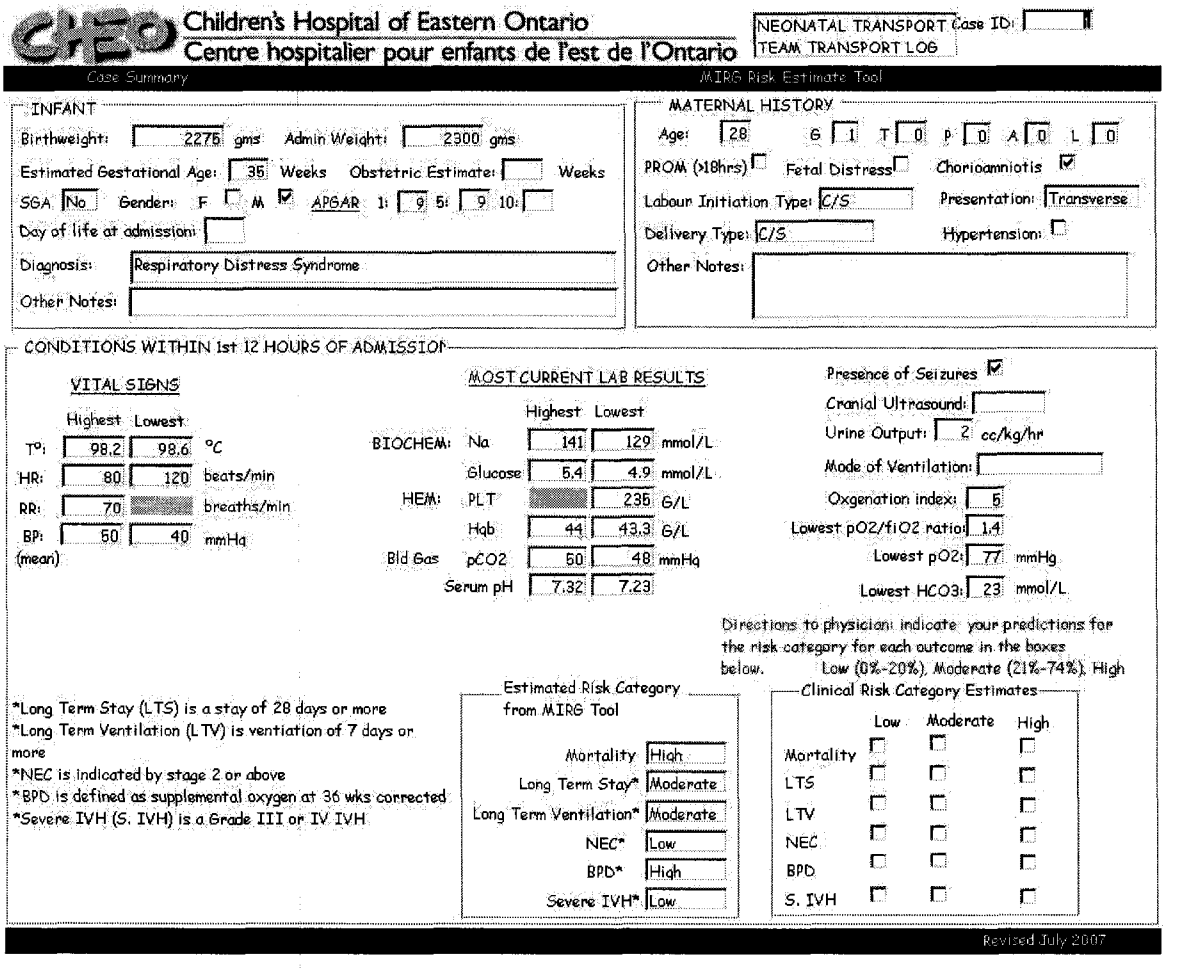

Figure 6.5. Interface used in the survey

Eighteen fields were added to allow physicians to select a checkbox corresponding to one of three estimated risk categories for each of the six outcomes. Most fields in the table were visible to the participants (mother's information, baby's information, SNAP variables, and risk estimates for certain groups of cases), while others were not displayed because they were not required or would interfere with the goals of the study (caselink and actual outcomes).

The second step, conducting the survey, was completed at CHEO. Interested physicians were provided with an envelope with one informed consent, one set of twenty cases to assess and one questionnaire. Participants who agreed with the study signed the informed consent and provided their predictions for twenty cases. Afterwards, they completed the questionnaire and returned the sealed envelope to the designated CHEO contact.

After conclusion of the on-site segment of this work, the results were analyzed. Four fields were extracted from the MS Access table (step (7) in Figure 6.4): the caselink, actual outcomes, estimated risk category and physicians' estimates. These were used in the quantitative analysis. The 120 cases were sorted into two groups (risk estimates provided to participants or not), and the physicians' risk estimates were compared to the actual outcomes and to the estimated risk category. The informed consent forms were sealed and archived, and answers from the questionnaires were compiled. The results of the survey are presented in the following section.

6.5 Discussion and Analysis of Study Results

The fifth objective of this thesis was:

> Compare clinicians' judgments to the CDS system in a clinical environment and analyze results to identify areas of success and areas that need improvement. Validate the case data and prediction display by means of a questionnaire to assess physicians' attitudes towards the prototype.

The study results are discussed in three sections: the first, *Qualitative Analysis*, addresses the information provided by the questionnaires and the second, *Quantitative Analysis*, examines the physicians' risk estimates in comparison to actual outcomes and to the estimated risk stratification. The third section discusses the strengths and weaknesses of the developed CDS system.

6.5.1 Qualitative Analysis

At the time of the survey, neonatologists had an average of 6 years' experience. Only two participants currently use a personal digital assistant for personal communications, one of whom uses one for work related communication as well. These answers were surprising because such devices are meant to increase the speed of information relay. No information in the comments section of the questionnaire revealed why personal digital assistants were not used, though the absence of electronic patient charts in the unit may lessen the potential benefits of using such devices. All respondents said they would respond positively to having access to electronic patient charts, which suggests the possibility for a positive outlook towards well designed computerized CDS systems and increased portable communication device use in the future. In fact question 2 of the survey, which asked physicians about the relevance of electronic tools for care delivery in the NICU, received the most positive answers of all questions, showing strong physician interest.

The second part of the questionnaire used a 4 point scale to obtain physicians' attitudes on fifteen topics. For the analysis a score of 2 was assigned to the most positive answer and a score of -2 was assigned to the least positive answer. For concurrence with Qi [2005] a score of 0 was deemed to be the cut-off point above which the opinions were deemed to be positive, and below which attitudes were deemed to be negative. The average score was calculated and is shown in the rightmost column of table 6-27. For example in question 1, the average score is (1*2 + 5*1 + 0*(-1) +0*(-2)) / 6 = 1.17. In the five point scale, neutral answers are attributed the weighing of 0 and do not influence the average score. Using the 4 point scale ensures all opinions affect the average score.

Table 6-27. Responses to part 2 of the questionnaire

Statements Please circle 1-4 for each of the 15 statements.	Number of answers in each category				
	2 strongly agree	1 agree	-1 disagree	-2 strongly disagree	Avg.
1. The case presentation software displayed predicted outcomes clearly and was easy to use.	1	5	0	0	1.17
2. The use of electronic tools can be relevant to the delivery of paediatric care in the NICU.	3	3	0	0	1.50
3. The outcomes and complications predicted were relevant to my patient management task.	1	5	0	0	1.17
4. I would use an ANN tool to assist the diagnosis and treatment of patients if the tool's outcome predictions were generally similar to my own.	0	5	1	0	0.67
5. I would use the ANN tool to assist the diagnosis and treatment of patients if the tool's outcome predictions were significantly different from my own.	0	2	3	1	-0.50
6. If my estimation of risk of death for a patient were significantly different from the ANN tool's estimation, I would most likely accept my own estimate.	0	6	0	0	1.00
7. The information (ANN tool's outcome prediction) could help me estimate mortality in terms of short-term survival (until discharge), the likely length of time a ventilator would be needed or estimate the patient's length of stay in NICU.	0	6	0	0	1.00
8. The information could help estimate the risk of complications (e.g., NEC, BPD and IVH) in a patient.	0	5	0	1	0.50
9. I have some knowledge about artificial neural networks.	0	4	0	2	0.00
10. An ANN tool that predicts with high accuracy a) could speed up my decision-making process. b) could increase the quality of care I deliver. c) could increase the accuracy of my predictions.	 1 0 1	 5 6 5	 0 0 0	 0 0 0	 1.17 1 1.17
11. The quality of medical information found in software applications I currently use (e.g.: electronic patient charts, computerized physician order entry) generally meets my standards.	0	1	5	0	-0.67
12. I would use the ANN tool if I was satisfied with the validity of the outputs.	0	6	0	0	1.00
13. In the development of prediction models, performance benchmarks of 90% specificity and 50% sensitivity or the current standard, whichever is highest were used. I am satisfied with an output that complies with these criteria.	0	6	0	0	1.00
14. I know neonatologists who support clinical decision support systems.	0	2	2	2	-0.67
15. Clinical decision support systems could reduce the cost of health care delivery in my area of expertise.	0	5	1	0	0.67

The average score for question one was 1.17, which suggests that physicians responded positively to the re-design of the interface, possibly due to the familiarity with the variable display and increased clarity in presentation of the outcomes.

The difference between statements 4 and 5 is that the first assumes the estimated risk stratification is similar to physicians' predictions while the second supposes that the risk estimates significantly differ from physicians' predictions. In both instances physicians stated they would more likely rely on their own estimates. The average score was 0.67 for the first question and -0.50 for the second. Thus the estimates seemed to be accepted as confirmation of a physician's diagnosis, perhaps to help reduce uncertainty when predictions are similar, and could be perceived as a colleague's opinion. This endorses the tool's role in decision support. If the estimated risk was significantly different, 4 out of 6 physicians would not likely use the CDS system to assist them in outcome estimation. This echoes findings in Qi's survey where 50% of attitudes to a similar statement were neutral, one was somewhat positive and two were strongly negative [2005]. As assumed, sceptical physicians opted for a weak negative answer when no neutral choice was possible. From the answers, we can see that some physicians are not interested in the system when the estimated risk category significantly differs from theirs.

When asking if the information provided by the estimated risk stratification would help in estimating particular outcomes such as mortality, duration of ventilation and the length of stay all physicians answered positively. Question 8 confirmed that the estimated risk models could also help the physicians estimate the risk of a patient developing any of the three modeled complications. The average score of 0.50 revealed that most physicians viewed this possibility positively, whereas one physician strongly disagreed. Most strongly negative answers came from this same physician.

Participants responded positively to the three parts of question 10: physicians expect that a CDS system that predicts with high accuracy could be helpful in three areas:

- the speed of decisions (in estimating the chance of outcomes)
- the quality of care delivered (especially in terms of parent counselling)
- the accuracy of predictions

Those topics are explicit areas of focus for CDS systems in clinical applications.

Question 10's results present a dramatic difference to the 2005 pilot survey where the ANN tool was voted 'unlikely' to speed up the decision making process [Qi 2005]. These results also suggest that doctors would be very accepting of well-designed CDS systems, though they are unhappy with the quality of current applications.

In summary, positive answers were found in questions regarding presentation, the relevance of electronic tools in the NICU, the outcomes and complications selected being relevant and the use of estimated risk stratification. The participants generally agreed that CDS systems can be helpful for their tasks as diagnosticians as well as having the potential to reduce some of the cost of healthcare delivery in their area of specialization.

In addition to statements 5 and 11, two statements elicited neutral or negative responses: knowledge about ANNs and knowing a fellow neonatologist advocating for CDS systems. There is a clear need to increase awareness about methods used by engineers to model physiologic processes since physicians are far more familiar with statistical methods (i.e.: logistic regression) than complex non-linear methods [Baxt 2004]. The best way to transmit technical details and potential benefits of a decision support system to physicians could be studied from a cognitive engineering point of view as physicians would likely be interested in information beyond the mathematical performance of ANNs (such as the validation of a models' plausibility by a panel of experts [Dreseitl & Ohno-Machado 2002]). Increased knowledge could also affect their likelihood to accept risk estimates as valuable information.

All the variables suggested by physicians in the first study and available within the first 12 hours were included in the prediction interface of this survey. There was no mention of missing variables in the comments; therefore it is believed that comments from physicians during the pilot survey were successfully addressed. Unfortunately two physicians discovered that the labels for two variables related to red blood cell counts, haemoglobin and hematocrit, were incorrectly assigned.

The third part of the questionnaire asked the participants to give their opinions on five topics to help guide future work. All participants liked having a percentage associated with each risk category, which suggests that the estimated risk models were

better received than a binary output. This assumption, based on prior knowledge, was the underling motivation for Zhou's [2006] work with the maximum likelihood estimation.

Answers to question 17 confirmed that the likelihood of long term disability is a desirable outcome to include in future versions of the tool, especially "for the counselling of parents". Answers to question 18 added the outcomes of cerebral palsy, blindness, deafness and a resulting intellectual quotient inferior to two standard deviations below the expected mean, to the list of possible future models. As increasing numbers of very low gestational age babies survive, the focus of desired models is expanding to include long term morbidities which are some of parents' biggest concerns when discussing treatment options with physicians [Meadow et al. 2004].

The last question asked participants for comments about any aspect of the study they felt was important. Some comments were unexpected, for example one participant did not like having the predictions available for only ten out of twenty cases and felt the survey was too short, while a second participant stated outcome predictions would only be useful for long-term outcomes. This last comment is surprising as it has been stated numerous times in literature that doctors, parents and hospital administrators desire more information about possible long and short-term outcomes [Meadow et al, 2004], [Stevens et al. 1994], [Zernikow et al. 1998], and one participant wrote that predictions for short term survival (48-72 hours) would be very useful.

Four participants stated that decision support in the format provided by the estimated risk stratification models could be of value for parent counselling, again supporting the change from a binary outcome to a categorized risk format. When asked about the possible use of case-matching, 4 out of 6 physicians seemed interested, which introduces the possibility of integrating case-matching to the estimated risk stratification models in future work. This feature could allow "instant recall of any case from thousands to be retrieved" [Frize & Walker 2000] and complements the estimated risk stratification models by providing a case-based explanation of the classification results in a way that ANNs do not.

6.5.2 Quantitative Analysis

The quantitative analysis begins with a comparison of physicians' estimates to actual outcomes, using estimates provided in the first ten cases in each physician's survey (without the estimated risk stratification). The small number of neonatologists at CHEO limits the depth of the statistical analysis; however some general remarks are made.

Having three risk categories of unequal breadth, compared to two equal classes in the previous survey, makes the analysis more difficult because the CCR is not easily determined (i.e. is a mortality classified as *Moderate Risk* a misclassification or not?). Table 6-28 compares physicians' estimates to actual results to show physicians' discrimination abilities.

Table 6-28. Physicians' predictions for cases without estimated risk stratification

		Risk Category		
		High	Moderate	Low
Mortality	Positives (n =11) (18.33%)	n = 0, 0%	n = 6, 55 %	n = 5, 45%
	Negatives (n=49) (81.67%)	n = 1, 2%	n = 10, 20%	n = 38, 78%
LOS	Positives (n =21) (35%)	n = 13, 62%	n = 2, 9%	n = 6, 29%
	Negatives (n=39) (65%)	n = 15, 38%	n = 4, 10%	n = 20, 52%
DOV	Positives (n =27) (45%)	n = 11, 40%	n = 8, 30%	n = 8, 30%
	Negatives (n=33) (55%)	n = 5, 15%	n = 10, 30%	n = 18, 55%
BPD	Positives (n =2) (3.33%)	n = 0, 0%	n = 2, 100%	n = 0, 0%
	Negatives (n=58) (96.67%)	n = 2, 4%	n = 14, 24%	n = 42, 72%
IVH	Positives (n =3) (5%)	n = 0, 0%	n = 2, 67%	n = 1, 33%
	Negatives (n=57) (95%)	n = 0, 0%	n = 8, 14%	n = 49, 86%
NEC	Positives (n =7) (11.67%)	n = 0, 0%	n = 0, 0%	n = 7, 100%
	Negatives (n=53) (83.33%)	n = 0, 0%	n = 7, 13%	n = 46, 87%

Averaged over the 60 cases, 32 % of rare outcomes (positives) were classified as High risk, and 73 % of common outcomes (negatives) were classified as Low risk. Physicians misclassified positives as Low risk and negatives as High risk in 9 % of instances, which is significant considering the prevalence of certain complications in the general NICU population is less than 3 %.

The estimated risk stratification was provided to physicians for the second group of ten cases they assessed, and their predictions are presented in table 6-29.

Table 6-29. Physicians' risk estimates compared to the estimated risk stratification

		PHYSICIAN ESTIMATES			**ESTIMATED RISK STRATIFICATION**		
		Risk Category			Risk Category		
		High	Moderate	Low	High	Moderate	Low
Mortality	Positives (n =11)	$n = 1$	$n = 6$	$n = 4$	$n = 2$	$n = 5$	$n = 4$
	Negatives (n=49)	$n = 0$	$n = 12$	$n = 37$	$n = 1$	$n = 5$	$n = 43$
LOS	Positives (n =21)	$n = 14$	$n = 2$	$n = 5$	$n = 8$	$n = 11$	$n = 2$
	Negatives (n=39)	$n = 19$	$n = 0$	$n = 20$	$n = 12$	$n = 10$	$n = 17$
DOV	Positives (n =27)	$n = 14$	$n = 3$	$n = 10$	$n = 5$	$n = 13$	$n = 9$
	Negatives (n=33)	$n = 2$	$n = 9$	$n = 22$	$n = 1$	$n = 10$	$n = 22$
BPD	Positives (n =2)	$n = 0$	$n = 0$	$n = 2$	$n = 0$	$n = 2$	$n = 0$
	Negatives (n=58)	$n = 4$	$n = 18$	$n = 36$	$n = 0$	$n = 21$	$n = 37$
IVH	Positives (n =3)	$n = 0$	$n = 2$	$n = 1$	$n = 0$	$n = 3$	$n = 0$
	Negatives (n=57)	$n = 0$	$n = 2$	$n = 55$	$n = 0$	$n = 3$	$n = 54$
NEC	Positives (n =7) =	$n = 0$	$n = 1$	$n = 6$	$n = 0$	$n = 4$	$n = 3$
	Negatives (n=53)	$n = 0$	$n = 10$	$n = 43$	$n = 0$	$n = 3$	$n = 50$

Physicians' discrimination ability on these cases resulted in 41 % of positive outcomes identified as high risk, and 81 % of negative outcomes classified as low risk, compared to 21 % and 77 % for the estimated risk stratification. According to these values, physicians' sensitivity is 20 % above that of the estimated risk stratification because a large number of "LOS" and "DOV" outcomes were correctly classified. However, the CCR can be misleading when outcomes are of low prevalence, which is why the results are presented in a different format in table 6-30 of the following section.

When provided with risk estimates, physicians misclassified 9 % of all outcomes (positives classified as Low risk and negatives classified as High risk), which is

consistent with the observations of table 6-27. The estimated risk stratification models faired somewhat better with 5 % of misclassifications.

This study was administered to too few participants to extrapolate the effect of providing risk estimates to all potential users of a system; however there was a 4% increase in correct classifications when the risk stratification was presented. A case by case comparison of physicians to ML ANN models was not performed because individual predictions cannot be considered representative of a group of physicians. Such a comparison may have been possible if the same 60 cases had been assessed by all participants. This arrangement will be considered in subsequent studies involving more tertiary care centres.

6.5.3 Discussion

The process that determines physicians' risk estimates for newborns are entirely different from the computations of a non-linear decision support system such as the one developed for this study, however in many instances the predicted risk categories were quite similar. Table 6-30 rates the discrimination ability of the study participants and models on two criteria: the highest number of positives classified as High risk, and the highest number of negatives classified as Low risk, for each outcome.

Table 6-30. Strengths of physicians and estimated risk stratification models during classification

	Most High Risk Positives		Most Low Risk Negatives	
	Physicians	Estimated Risk Stratification	Physicians	Estimated Risk Stratification
Mortality		√		√
LOS ≥28 days	√		√	
DOV ≥8 days	√		√	√
BPD		√		√
IVH		√	√	
NEC		√		√

The physicians were very successful at determining which cases required short or extended stays in the NICU and short or extended ventilation periods. For the latter outcome, the estimated risk stratification correctly classified low risk cases as well as the

physicians, which also was the case in Qi's study which used 24 hours as the ventilation criteria. For the complication of IVH, physicians correctly classified more negative cases as low risk. This is surprising given the very high specificity (≥99%) of the estimated risk stratification model. This may show that physicians are very proficient at determining when a patient is at low risk for certain rare complications, or at ruling out certain possibilities.

There are some cases which are practically impossible to predict correctly when using data available during the first 12 hours after admission. For example nosocomial infections, which by definition are not characterized by day 1 data, are impossible to predict for individuals, though group prediction models are likely feasible. It is uncertain which percentage of cases in the CNN and CHEO databases were affected by this type of secondary illness.

Estimating risk with CDS Systems

It is unclear whether physicians considered the time spent assessing the validity of the estimated risk categories, as they would a colleague's opinion, in the usefulness or likelihood to use the CDS system. This should be determined in future work because it is known that the time that a CDS system adds to a patient consultation affects a physician's intentions to use it. For example, in a study of an Internet health application which aimed to help physicians to determine if a patient requires a referral to a secondary care center, 70 % of doctors intended to use the system if it lengthened patient consultations by 2 minutes compared to only 23 % if the time added was 5 minutes [Van Schaik et al. 2004].

The average time required for physicians to provide predictions for the six outcomes on 20 cases and complete the questionnaire was 34 minutes for the study in this thesis. In this respect, the estimated risk stratification models offer only a minimal temporal advantage, however they are unaffected by the stress, fatigue or work overload issues that could affect diagnosticians.

Techniques of machine learning such as ANNs can extract useful information and present knowledge in a way that is coherent with physicians' expectations. For all outcomes considered in this thesis, the best structure in all but one case included hidden

nodes. This ability to model non-linear relationships is one reason that ANNs are widely used in medical applications [Baxt1994], [Dreiseitl & Ohno-Machado 2002].

Although ANNs have the possibility to help physicians with their decision-making, there are limits to their application. The ANN models typically do not perform well on un-seen cases which differ from the testing and training cases. This is an important consideration knowing that each NICU can have different mortality and morbidity rates, which may result from different population characteristics [Frize & Walker 2000], [Lee 2006].

Physicians' clinical experience and knowledge is limited by "time, location, demographics, budgets, and capacity" [Qi 2005], while ANNs can incorporate to model more cases than any physician could assess and remember [Frize & Walker 2000]. Another advantage of using ANNs is that models can be re-trained as new data becomes available, a process which incorporates new knowledge into the models. Unfortunately it takes time to accumulate a sufficient number of cases to train a generalizable model, compared to physicians who can adapt to new knowledge or procedures very quickly.

There is a difficulty in assessing physician's predictions which is caused by using retrospective case data in lieu of current patients. Requiring an abstractor to extract certain variables and a presentation interface to display the data may affect physicians' discrimination ability.

Strengths and Limitations of Physicians

As indicated by some comments in this study's questionnaires, physicians have a great ability to detect incorrect values whereas the MSE and ML ANNs used in this work cannot. Data integrity is always important, whether variables are abstracted individually or automatically archived.

Physicians are very well equipped to provide accurate diagnoses and prognoses under a wide array of conditions, and are able to adapt the significance of observations to individual cases. There is also a level of trust that exists between doctors and their patients that cannot be provided by correct classification rates. This relationship is especially important in the NICU when there is uncertainty regarding possible outcomes and decision-making is shared with parents [Whitney 2003].

This study acquired physicians' attitudes towards an interface displaying the estimated risk category for several important outcomes relevant to care in the NICU. Physicians' performance was also compared to the models, which yielded information that can be used to guide the refinement of current models. The extent of this study was in part limited by the number of and accessibility to neonatologists and to simplify the analysis, others who may potentially benefit such as hospital administrators and parents were not surveyed.

"To be useful, the ANN tool should be able to predict non-survivors with at least as much accuracy as the physician" [Walker et al. 2002]. For this small collection of 60 cases, the models did in fact classify more mortalities as High risk and more or as many survivors as Low risk than did the physicians regardless of whether they were provided with risk estimates (table 6-29) or not (table 6-28).

Comparing doctor's estimates to SNAP variable models "provides insight into clinical decision making and has important applications in improving direct patient care and appropriate allocation of medical resources and medical training", [Stevens et al. 1994]. This study helped identify cases which are successfully predicted by physicians and areas that can be improved upon in future models, such as the sensitivity and general discrimination ability of complication models.

CHAPTER 7 CONCLUSIONS

7.1 Conclusions

This research involved the development of prediction models and case presentation software as well as the validation of models with physicians, which are important steps in the work towards clinical implementation of the Medical Information Technology Research Group's clinical decision support system. Important outcomes identified by MIRG's medical partners were modeled using data from the first 12 hours of admission, and the estimated risk stratification was obtained and incorporated into a visual display. By using engineering techniques to develop the risk estimates and validating models using doctor's predictions, this thesis helps move MIRG's research projects towards full clinical integration.

Two classification ANNs were applied to the imputed database: a MSE ANN which improved the sensitivity of models by removing variables, and a ML ANN which created an estimated risk stratification for all outcomes.

7.2 Contributions to Knowledge

Several objectives were met during this research and provided valuable information that complements the long term goals of MIRG's research in clinical decision support.

1. It was determined that an imputation with uniform weights in the k-NN allowed for the determination of a mortality model that performed better than the SNAPPE-II and the SNAP models.
2. Minimal datasets for important clinical outcomes and complications were determined with the MSE ANN, from which risk estimate models were obtained using the ML estimation.
3. A study of the Informed Consent process for the ethical process of the clinical study resulted in a publication at the 28th International Conference of the IEEE EMBS in NY, NY.
4. A study protocol and questionnaires that can be used as a guide in future trials were created.

5. Answers from questionnaires determined that physicians support the use of an accurate CDS system to:
    a. Increase the speed of decisions (in estimating the chance of outcomes)
    b. Improve the quality of care delivered (especially in terms of parent counselling)
    c. Improve the accuracy of predictions
6. An analysis of physicians' discrimination ability revealed that physicians were very accurate in their identification of cases at high and low risk for extended ventilation and extended length of stay, while their correct classification rate was below that of the estimated risk stratification models.

7.3 Future Work

7.3.1 Model Refinement

The NTISS score provides information on the level of therapeutic intervention delivered and as such is related to the severity of illness. It would be interesting to see if the inclusion of some of the NTISS variables to the minimal variable sets of SNAP variables could improve the prediction accuracy for certain outcomes.

The categorization of duration of ventilation into week-long periods has been accomplished by MIRG members in adult models of post-operative outcomes [Buskard et al. 1994]. The concept could be extended to categorize lengths of stay in the NICU to provide more specific risk estimates to physicians, hospital administrators and parents.

The number of NICU patients risking nosocomial infection is increasing "because of the improved survival of very low birth weight infants and their need for invasive monitoring" [Adams-Chapman & Stoll 2002], and the EPIC abstractor's manual states a varying infection rate (6-40%) between NICUs in Canada [Lee 2006]. Nosocomial infections have been studied to identify potential vectors and risk factors leading to infection, however no evidence was found in PubMed to suggest that they have previously been studied with non-linear networks.

Previous attempts at using a k-NN for outcome prediction, as opposed to variable imputation, were unsuccessful when using uniform and outcome-specific weights. A

new direction would be to devise an improved similarity calculation to present the most closely matched cases (i.e.: 5 survivors and 5 non-survivors). Finally, adding CBR ability to interface, and determining if the combination of estimated risk stratification and similar case display helps increase the accuracy of predictions.

7.3.2 Human-Computer Interaction Research

The final CDS system's interface is left as future work as it will require extensive cognitive engineering, network security and medical knowledge to properly address the characterization of user needs and the improvement of professional-computer interaction [Tang & Patel 1994].

As part of the work of this thesis, ethical approval was obtained for a separate study at CHEO to observe the effect of an evidence-based CDS tool for physicians and parents, on the parents' decision-making. The concept is established; however the implementation remains for a future student as an entire project so that it may be probed in more depth than would have been possible in this thesis.

7.3.3 System Integration

With the new knowledge about the prediction abilities and attitudes of physicians towards the CDS system obtained in this clinical survey, future work may focus on the refinement and integration of developed components into the proposed 'Semantics Web Services for healthcare framework' suggested by Catley [2007].

It is important to establish a plan for clinical integration including reading best practices, regulations and drafting a clinical integration protocol with medical partners to ensure the compatibility of all future work with the long-term goals of MIRG researchers and physician users.

Lastly, a comprehensive study of the real-time data in the repository could be conducted to determine trends in missing values and identify potential areas for improvement (i.e. differentiating between a ventilator that is turned off and one whose data line has been disconnected or is associated with the incorrect incubator).

REFERENCES

1. [Alberdi et al. 2000]
Alberdi E, Gilhooly K, Hunter J, Logie R, Lyon A, McIntosh N, Reiss J. "Computerisation and decision making in neonatal intensive care: a cognitive engineering investigation". Journal of Clinical Monitoring and Computing, 16:85-94 2000.

2. [Ammenwerth et al. 2003]
Ammenwerth E, Kaiser F, Wilhelmy I. Höfer S. Evaluation of user acceptance of information systems in health care – the value of questionnaires -. Proceedings of Medical Informatics Europe, 4 – 7 May 2003, St. Malo, France. Pages 643-648.

3. [Annibale & Hill 2006]
Annibale DJ., Hill J. Periventricular Hemorrhage - Intraventricular Hemorrhage. Emedicine from WebMD, updated June 19th 2006.

4. [Bariciak, 2006]
Dr. E. Bariciak, Neonatologist at CHEO, personal communication, December 8th 2006.

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

6. [Berner 1998]
Berner E. Clinical decision support systems theory and practice. Alabama, United States of America, Springer 1998, 292p.

7. [Bishop 1995]
Bishop C. Neural networks for pattern recognition. Oxford, [England]: Clarendon Press, 1995. 482p.

8. [Buskard et al. 1994]
Buskard T, Stevenson M, Frize M, and Solven FG. "Estimation of ventilation, length of stay and mortality using artificial neural networks," in Proc. Canadian Conf. on Electrical and Computer Engineering, vol. 2, 1994, pp. 726-729

9. [Catley et al. 2005]
Catley C, Frize M, Walker RC, Petriu DC. Predicting preterm birth using artificial neural networks. 18th IEEE Symposium on Computer-Based Medical Systems, 2005. Proceedings. (2005) : 103- 108.

10. [Catley 2007]
Catley C. "An integrated hybrid data mining system for preterm birth risk assessment based on a 'semantic web services for healthcare' framework". Phd Thesis, Carleton University, 2007, 252p.

11. [Catley & Frize 2003]
Catley C and Frize M. "A Prototype XML-Based Implementation of an Integrated 'Intelligent' Neonatal Intensive Care Unit" Proc of the 4th IEEE Conference on Information Technology Applications in Biomedicine (2003): 323-25.

12. [CNN 2004]
Canadian Neonatal Network, "Abstractor's Manual", July 2004, 98p., www.canadianneonatalnetwork.org/annual.shtml

13. [Chien et al. 2002]
Chien L-Y, Whyte R, Thiessen P, Walker R, Brabyn D, Lee SK. SNAP-II Predicts Invtraventricular Hemorrhave and Chronic Lung Disease in the Neonatal Intensive Care Unit. Journal of Perinatology, January 2002, Vol. 22 No. 1 pp. 26-30.

14. [Chismar & Wiley-Patton 2003]
Chismar WG,Wiley-Patton S. Does the extended technology acceptance model apply to physicians. System Sciences, 2003. Proceedings of the 36th Annual Hawaii International Conference on 6-9 Jan 2003 Page(s):8 pp.

15. [Courtright et al. 2001]
Courtright CG, Crawford RS. Klubert DM. Criteria for developing clinical decision support systems. 14th IEEE Symposium on Computer-Based Medical Systems, 2001. CBMS 2001. Proceedings. 26-27 July 2001 Page(s):270 - 275

16. [Ennett 2003]
Ennett CM. Imputation of Missing Values by Integrating Artificial Neural Networks and Case-based Reasoning, PhD Thesis. Department of Systems and Computer Engineering, Carleton University, Ottawa Ontario, 2003.

17. [Ennett et al. 1999]
Ennett CM, Frize, M.; Shaw RE. Methodologies for predicting coronary surgery outcomes. BMES/EMBS Conference, 1999. Proceedings of the First Joint Volume 2, 13-16 Oct. 1999 Page(s):1240 vol.2

18. [Ennett et al. 2001]
Ennett CM, Frize M, Walker CR. Influence of missing values on artificial neural network performance. Medinfo. 2001;10 (Pt 1):449-53.

19. [Ennett et al. 2003]
Ennett CM, Frize M, Scales N. Evaluation of the logarithmic-sensitivity index as a neural network stopping criterion for rare outcomes. Proc of the 4th Annual IEEE Conf on Information Technology Applications in Biomedicine, UK 2003.

20. [Ennett & Frize 1998]

Ennett CM, Frize M. An investigation into the strengths and limitations of artificial neural networks: an application to an adult ICU patient database. Proc AMIA Symp 1998:998.

21. [Ennett & Frize 2003]
Ennett CM, Frize M. Weight-elimination neural networks applied to coronary surgery mortality prediction. Information Technology in Biomedicine, IEEE Transactions on, Volume: 7, Issue: 2, June 2003. page(s): 86- 92

22. [Ewing et al. 2003]
Ewing G, Freer Y, Logie R, Hunter J, McIntosh N, Rudkin S, Ferguson L. Role and experience determine decision support interface requirements in a neonatal intensive care environment. J Biomed Inform. 2003 Aug-Oct;36(4-5):240-9.

23. [Finster & Wood 2005]
Finster M, Wood M. The Apgar Score Has Survived the Test of Time. Anesthesiology 2005; 102:855–7.

24. [Fishbein & Ajzen 1975]
Fishbein M, Ajzen I. Belief, attitude, intention, and behavior. Reading, MA: Addison-Wesley. (1975).

25. [Frize et al. 1998]
Frize M, Wang L, Ennett CM, Nickerson BG, Solven FG, Stevenson M. New Advances and Validation of Knowledge Management Tools for Critical Care Using Classifier Techniques. Proc AMIA Annual Symposium, 1998 - amia.org.

26. [Frize, Ennett & Charette 2001]
Frize M, Ennett CM, Charette E. "Automated optimization of the performance of artificial neural networks to estimate medical outcomes." Proc IEEE ITAB-ITIS 2000:168-73.

27. [Frize et al. 2003]
Frize M, Walker RC, Ennett CM. Development of an Evidence-Based Ethical Decision-Making Tool for Neonatal Intensive Care Medicine. Proceedings of the 25th Annual International Conference of the IEEE EMBS Cancun, Mexico. September 17-21, 2003.

28. [Frize et al. 2004]
Frize M, Catley C, Walker CR, Petriu DC, Yang L. Towards a Web Services Infrastructure for Perinatal, Obstetrical, and Neonatal Clinical Decision Support International Conference of the Engineering in Medicine and Biology Society, 2004. EMBC 2004, Conference Proceedings.26th Annual. Volume 2, 2004 Page(s):3334 - 3337

29. [Frize & Walker 2000]
Frize M, Walker CR. Clinical decision-support systems for intensive care units using case-based reasoning. Medical Engineering & Physics 22 (2000) 671-677.

30. [Frize et al. 2005]
Frize M, Yang L, Walker RC, O'Connor AN. Conceptual Framework of Knowledge Management for Ethical Decision-Making Support in Neonatal Intensive Care. IEEE Transactions on Information Technology in Biomedicine, Vol.9, NO.2, June 2005 p. 205-215.

31. [Fu 1994]
Fu L. Neural Networks in Computer Intelligence. United States of America, McGraw-Hill. 1994. 460p.

32. [Goh 1995]
Goh TC. Back-propagation neural networks for modeling complex systems. Artificial Intelligence in Engineering, 9(1995), pp.143-151.

33. [Jerebko et al. 2003].
Jerebko AK, Malley JD, Franaszek M, Summers RM. (2003). Multiple neural network classification scheme for detection of colonic polyps in CT colonography data sets. Academic Radiology, 10, 154-160.

34. [Joseph et al., 1998]
Joseph KS, Kramer MS, Marcroux S, Ohlsson A, Wen SW, Allen A, Platt R. "Determinants of preterm birth rates in Canada from 1981 through 1983 and from 1992 through 1994". The New England Journal of Medicine. November 12, 1998; pp. 1434-1439.

35. [Kaplan 2001]
Kaplan B. Evaluating informatics applications--clinical decision support systems literature review. Int J Med Inform. 2001 Nov; 64(1):15-37.

36. [Khalil et al. 1995]
Khalil KA, El-Amrawy SM, Ibrahim AG, El-Zeiny NA and Greiw AE. Pattern of growth and development of premature children at the age of two and three years in Alexandria, Egypt (Part II). Eastern Mediterranean Health Journal. Volume 1, Issue 2, 1995, Page 186-193. http://www.emro.who.int/Publications/EMHJ/0102/03.htm

37. [Kramer et al. 2001]
Kramer MS, Platt RW, Wen SW, Joseph KS, Allen A, Abrahamowicz M, Blondel B, Breart G. Fetal/Infant Health Study Group of the Canadian Perinatal Surveillance System. "A new and improved population-based Canadian reference for birth weight for gestational age". Pediatrics. 2001 Aug;108(2):E35.

38. [Larcher & Hird 2002]
Larcher V, Hird MF. Withholding and withdrawing neonatal intensive care. Current Paediatrics (2002) 12, 470-475.

39. [Livingstone & Manallack 1993]
Livingstone DJ. Manallack DT. "Statistics using neural networks chance effects," J Med Chem. 1993; 36: 295-97.

40. [Lee 2006]
Lee K. "EPIC Evidence-based Practice Identification and Change". EPIC/PHSI Training Workshop November 9 & 10, 2006, Toronto ON.

41. [Masters 1993]
Masters T. Practical Neural Network Recipes in C++. San Diego, CA: Academic Press, 1993 493p.

42. [McLaughlin et al. 1999]
McLaughlin F, Rusen ID and Liu SL. Canadian Perinatal Surveillance System, Public Health Agency of Canada, October 1999 http://www.phac-aspc.gc.ca/rhs-ssg/factshts/pterm_e.html.

43. [Meadow et al. 2004]
Meadow W, Lee G, Lin K, Lantos J. Changes in mortality for extremely low birth weight infants in the 1990s: implications for treatment decisions and resource use. Pediatrics. 2004 May;113(5):1223-9.

44. [Noqueira et al. 2007]
Nogueira BM, Santos TRA and Zárate LE. Comparison of Classifiers Efficiency on Missing Values Recovering: Application in a Marketing Database with Massive Missing Data. Proceedings of the 2007 IEEE Symposium on Computational Intelligence and Data Mining (CIDM 2007), pp.66-72.

45. [O'Connor et al. 2002]
O'Connor AM, Jacobsen MG, Stacey D. An Evidence-Based Approach to Managing Women's Decisional Conflict. JOGNN Clinical Issues, 31, 570-581; 2002.

46. [Olden & Jackson, 2002]
Olden J, Jackson D. "Illuminating the 'black box': a randomization approach for understanding variable contributions in artificial neural networks," Department of Zoology, University of Toronto, Ecological Modelling, 2002; 154:135–150.

47. [Penny & Frost 1996]
Penny W, Frost D. Neural networks in clinical medicine. Med Decis Making 1996:16:386-398.

48. [Pyper 2007]
Pyper C. "Final Report for the Pediatric Resident Decision Support System", 4th Year Final Report. University of Carleton, May 2007, 81p.

49. [Qi 2005]
Qi L. Evaluation of an Artificial Neural Network Tool for Neonatal Intensive Care Units. MASc thesis. Department of Systems and Computer Engineering, Carleton University, Ottawa Ontario.

50. [Richardson et al. 1993]
Richardson DK, Gray JE, McCormick MC, Workman K, Goldmann DA. Score for neonatal acute physiology: A physiologic severity index for neonatal intensive care. Pediatrics. Evanston: Mar 1993.Vol.91, Iss. 3; pg. 617.

51. [Richardson et al. 2001]
Richardson, DK, Corcoran JD, Escobar GJ, Lee SK. SNAP-II and SNAPPE-II: Simplified newborn illness severity and mortality risk scores. Journal of Pediatrics Vol: 138, Issue: 1, January, 2001 pp. 92-100.

52. [Rybchynski 2005]
Rybchynski D. Design of an Artificial Neural Network Research Framework to Enhance the Development of Clinical Prediction Models. MASc Thesis. School of Information Technology and Engineering, University of Ottawa, Ottawa Ontario 2005.

53. [Safran 2003]
Safran C. The collaborative edge: patient empowerment for vulnerable populations. International Journals of Medical Informatics 69 (2003) 185-190.

54. [Shankaran et al. 2004]
Shankaran K, Johnson Y, Langer JC, Vohr BR, Fanaroff AA, Wright LL, Poole WK. Outcome of extremely-low-birth-weight infants at highest risk: gestational age < or =24 weeks, birth weight < or =750 g, and 1-minute Apgar < or =3. Am J Obstet Gynecol. 2004 Oct;191(4):1084-91.

55. [Stevens et al. 1994]
Stevens SM, Richardson DK, Gray JE, Goldman DA McCormick MC. Estimating neonatal mortality risk: an analysis of clinicians' judgements. Pediatrics 1994;93(6):945-950.

56. [Tang & Patel 1994]
Tang PC, Patel VL. Major issues in user interface design for health processional workstations: summary and recommendations. Int. J. Biomed. Comput. 34 (1994) 139-148.

57. [Townsend 2006]
Townsend, D. "Informed Consent in Biomedical Engineering". Proceedings of the 28th International Conference of the IEEE EMBS in NY, NY. August 30th-September 3rd 2006.

58. [Trigg 1997]
Trigg HCE. "An investigation of methods to enhance the performance of artificial neural networks used to estimate ICU outcomes." Master's thesis, Department of Electrical Engineering, University of New Brunswick, Fredericton, NB, Jan 1997.

59. [Troyanskaya et al. 2001]
Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D and Altman R. Missing value estimation methods for DNA microarrays. Bioinformatics Vol. 17 no. 6 2001, pp. 520-525.

60. [Van Riper 2001]
Van Riper M. Family-provider relationships and well-being in families with preterm Infants in the NICU. Heart Lung. 2001 Jan-Feb;30(1):74-84.

61. [Van Schiak et al. 2004]
Van Schiak P, Flynn D, Van Wersch A, Douglass A, Cann P. The acceptance of a computerised decision-support system in primary care: A preliminary investigation. Behaviour & information technology. 2004, vol. 23, n°5, pp. 321-326.

62. [Walker 2006] Personal conversation at CHEO, August 2006.

63. [Weigend et al. 1990]
Weigend AS, Rumelhart DE, Huberman BA. "Back-propagation, weight-elimination and time series prediction." In DS Tourestzky, JL Elman, TJ Sejnowski, GE Hinton, eds. Proc 1990 Connectionist Models Summer School. Morgan Kaufmann: San Mateo. 1990:105-116.

64. [Whitney 2003]
Whitney S. A New Model of Medical Decisions: Exploring the Limits of Shared Decision Making. Medical Decision Making. 23 (4): 275.

65. [Wu et al. 2007]
Wu Z, Li C, Kee-Yin Ng J and Leung K. Location Estimation via Support Vector Regression. IEEE Transactions on Mobile Computing. March 2007 (Vol. 6, No. 3) pp. 311-321.

66. [Zamboni 2007]
Zamboni I. "Risk Assessment of Necrotizing Enterocolits in NICU". Masters of Applied Science, Carleton University, 2007. [not yet published]

67. [Zernikow et al. 1998]
Zernikow B, Holtmannspoetter K, Michel E, Pielemeier W, Hornschuh F, Westermann A, Hennecke KH. Artificial neural network for risk assessment in preterm neonates. Arch Dis Child Fetal Neonatal Ed. 1998 Sep;79(2):F129-34.

68. [Zernikow et al. 1999]
Zernikow B, Holtmannspoetter K, Michel E, Hornschuh F, Groote K, Hennecke K-H. Predicting length-of-stay in preterm neonates. Eur J Pediatr (1999) 158:59-62.

69. [Zhou 2006]
Zhou D. Adding Probability to Neural Network Prediction of NICU Mortality. Masters of Science, University of Ottawa, May 1st 2006. 104p.

APPENDIX A – Scoring Systems

Table A-1. SNAP variable list and scoring system [Richardson et al. 1993]

Variable		Points scored			
		0	1	3	5
Mean blood pressure	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	30-50	<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 neutrophil	High	<0.21	≥0.21	–	–
Absolute neutrophil count	Low	>999	500-999	<500	–
Platelet count	Low	>100	30-100	0-29	–
Blood urea nitrogen	High	<40	40-80	>80	
Creatinine	High	<1.2	1.2-2.4	2.5-4.0	>4.0
Urine output	Low	>0.9	0.5-0.9	0.1-0.49	<0.1
Indirect bilirubin	High				
- for birth weight>2kg		<15	15-20	>20	–
- for birth weight<2kg		<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	–
Total calcium	High	<12	>12	–	–
	Low	>6.9	5.0-6.9	<5.0	–
Ionized calcium	High	<1.4	>1.4	–	–
	Low	>1.0	0.8-1.0	<0.8	–
Glucose	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	7.10-7.19	<7.10
Presence of seizures		None	Single	Multiple	–
Presence of apnea		none	Response to stimuli	No response to stimuli	Complete apnea
Stool guaiac		Negative	Positive	–	–

* Additional points scored for the Perinatal Extension (SNAPPE) are:

Birth weight ≤749 g	30 points
Birth weight 750-999 g	10 points
Apgar < 7 at 5 minutes	10 points
Small for gestational age (<5th percentile)	5 points

Table A-2. SNAP-II variable list and scoring system [Richardson et al. 2001]

Variable	Range	Points
Lowest blood pressure	MBP[a] 20-29 mmHg	9
	MBP[a] <20 mmHg	19
Lowest temperature	95-96 °F	8
	<95 °F	15
Lowest PO_2/FiO_2 ratio	1.0-2.49	5
	0.3-0.99	16
	<0.3	28
Lowest serum pH	7.10-7.19	7
	< 7.10	16
Multiple seizures	>1	19
Lowest urine output	0.1-0.9 mL/kg/h	5
	<0.1 mL/kg/h	18
Birth weight	750-999 g	10
	<750 g	17
Small for gestational age	< 3rd percentile	12
Apgar score at 5 minutes	< 7	18

[a] MBP= mean arterial blood pressure

Table A-3. Apgar scoring system [Finster & Wood 2005]

	0	1	2
Appearance	Pale	Blue	Pink
Pulse	Absent	<100	>100
Grimace	Absent	Grimace	Cry active
Activity	Limp	Some tone	Active
Respiration	Absent	Irregular	Regular and cry

Table A-4. Variables presented to physicians [Qi 2005]

Admission Information	**SNAP**	**Mother's information**
Case number	Blood pressure low	Mother's age
Gender	Respiratory rate	Gravida
Birth weight	Temperature	Total abortions
Gestational age (Paediatric estimate)	Serum pH	Labour initiation type
Gestational age (Obstetric estimate)	Lowest pO2-FiO2	Delivery type
Apgar Score @ 1, 5 and 10 minutes	Lowest pO2-MAWP	Maternal hypertension
Births this pregnancy	Seizures	Presentation
Admission weight	Urine output CC's	Prenatal care
		Antenatal corticosteroid
NTISS	**Diagnosis/procedures***	Diabetes
Supplemental O2	PDA	Chorioamnionitis
CPAP	RDS	Tocolysis
Mechanical ventilation	Pneumothorax	
Ventilation with relaxant		
High frequency ventilation		
Surfactant	*if diagnosed within first	
Nitric oxide	12 hours of admission	

APPENDIX B – Values for Control Parameters

Table B-1. Starting values for the ANN control parameters

Parameter	Start Value
No. of layers	3
Hidden nodes	0 :2n+1
Weight-elimination	Yes
RNG seed	5
Error ratio	1.02
Lambda	0.001
Lambda decrement	0.999
Lambda increment	1.001
Learning rate	0.00001
Learning rate decrement	0.999
Learning rate increment	1.001
Momentum	0.8
Weight scale factor	0.01

Table B-2. Starting values of control parameters for the maximum likelihood estimation

Parameter	Start Value
No. of layers	3
Hidden nodes	0 :2n+1
Weight-elimination	No
RNG seed	7
Error ratio	1.02
Lambda	0.001
Lambda decrement	0.999
Lambda increment	1.001
Learning rate	0.00005
Learning rate decrement	0.99
Learning rate increment	1.01
Momentum	0.4
Weight scale factor	0.01

APPENDIX C – Estimated Risk Stratification

Tables C-1 to C-5 show the estimated risk stratification result on the test set.

Table C-1. The estimated risk stratification of the length of stay model

Probability Group	Expected Positives	Observed Positives	Expected Negatives	Observed Negatives	Total Cases in Group
<0.1	89	98	2020	2011	2109
0.1-0.2	86.4	78	522.6	530	609
0.2-0.3	68.6	57	211.4	223	280
0.3-0.4	76.7	74	142.3	145	219
0.4-0.5	83	85	102	100	185
0.5-0.6	113.6	120	91.4	85	205
0.6-0.7	116	173	89	82	255
0.7-0.8	212	219	70	63	282
0.8-0.9	270.1	261	74.9	57	318
0.9-1.0	60.7	57	5.3	9	66

Table C-2. The estimated risk stratification of the ventilation model

Probability Group	Expected Positives	Observed Positives	Expected Negatives	Observed Negatives	Total Cases in Group
<0.1	75.9	81	3046.1	3041	3122
0.1-0.2	59.6	55	354.4	359	414
0.2-0.3	51.2	48	156.8	160	208
0.3-0.4	54.7	58	104.3	101	159
0.4-0.5	68.9	64	83.1	88	152
0.5-0.6	60.4	60	49.6	50	11
0.6-0.7	69.8	67	38.2	41	108
0.7-0.8	100	101	34	33	134
0.8-0.9	85.1	82	15.9	19	101
0.9-1.0	17.5	15	1.5	4	19

Table C-3. The estimated risk stratification of the BPD model

Probability Group	Expected Positives	Observed Positives	Expected Negatives	Observed Negatives	Total Cases in Group
<0.1	56.4	56	3531.6	3520	3558
0.1-0.2	41.2	37	242.8	247	284
0.2-0.3	42.1	50	126.9	119	169
0.3-0.4	51.5	47	95.5	100	147
0.4-0.5	68	79	84	73	125
0.5-0.6	56.2	56	46.8	47	103
0.6-0.7	41.3	38	22.7	26	64
0.7-1.0	15.9	12	5.1	9	21

Table C-4. The estimated risk stratification of the IVH model

Probability Group	Expected Positives	Observed Positives	Expected Negatives	Observed Negatives	Total Cases in Group
<0.1	81.1	70	3922.9	3924	4004
0.1-0.2	48.5	42	294.5	301	343
0.2-0.3	26.6	21	83.4	89	110
0.3-0.4	14.5	11	27.5	31	42
0.4-1.0	15	11	14	18	29

Table C-5. The estimated risk stratification of the NEC model

Probability Group	Expected Positives	Observed Positives	Expected Negatives	Observed Negatives	Total Cases in Group
<0.1	50.6	43	3752.3	3750	3793
0.1-0.2	35.8	29	267.5	266	295
0.2-0.3	23.4	15	142.8	142	157
0.3-0.4	15.9	7	53.6	54	61
0.4-0.5	14	5	41	42	47
0.5-1.0	16.3	7	15.2	18	25

APPENDIX D – Survey and Questionnaire

Artificial Neural Networks (ANN) Tool Survey & Questionnaire

Description:

Our research group is now developing a decision-support system for potential future clinical application. The tool uses artificial neural networks to estimate the risk of outcomes for the single newly admitted patient to NICU: mortality, the duration of artificial ventilation, the length of stay and the diagnosis of three important complications: bronchopulmonary dysplasia, intraventricular hemorrhage and necrotizing enterocolitis. This questionnaire is designed to collect information on the attitude of physicians towards using this tool and suggestions for its improvement.

In order to analyze the results more objectively, before starting the questionnaire, please answer some questions about yourself:

1. Approximately how many years have you worked as a neonatologist? _____

2. Do you use a personal digital assistant:

   ☐ for personal communication?

   ☐ for work related communication?

3. Would you respond positively to having access to electronic patient charts?

   ☐ yes ☐ no.

Please circle 1-4 for each of the 15 statements.

Statements	1 strongly agree	2 agree	3 disagree	4 strongly disagree
1. The case presentation software displayed predicted outcomes clearly and was easy to use.	1	2	3	4
2. The use of electronic tools can be relevant to the delivery of paediatric care in the NICU.	1	2	3	4
3. The outcomes and complications predicted were relevant to my patient management task.	1	2	3	4
4. I would use an ANN tool to assist the diagnosis and treatment of patients if the tool's outcome predictions were generally similar to my own.	1	2	3	4
5. I would use the ANN tool to assist the diagnosis and treatment of patients if the tool's outcome predictions were significantly different from my own.	1	2	3	4
6. If my estimation of risk of death for a patient were significantly different from the ANN tool's estimation, I would most likely accept my own estimate.	1	2	3	4
7. The information (ANN tool's outcome prediction) could help me estimate mortality in terms of short-term survival (until discharge), the likely length of time a ventilator would be needed or estimate the patient's length of stay in NICU.	1	2	3	4
8. The information could help estimate the risk of complications (e.g., NEC, BPD, IVH) in a patient.	1	2	3	4
9. I have some knowledge about artificial neural networks.	1	2	3	4
10. An ANN tool that predicts with high accuracy a) could speed up my decision-making process. b) could increase the quality of care I deliver. c) could increase the accuracy of my predictions.	 1 1 1	 2 2 2	 3 3 3	 4 4 4
11. The quality of medical information found in software applications I currently use (e.g.: electronic patient charts, computerized physician order entry) generally meets my standards.	1	2	3	4
12. I would use the ANN tool if I was satisfied with the validity of the outputs.	1	2	3	4
13. In the development of prediction models, performance benchmarks of 90% specificity and 50% sensitivity or the current standard, whichever is highest were used. I am satisfied with an output that complies with these criteria.	1	2	3	4
14. I know neonatologists who support clinical decision support systems.	1	2	3	4
15. Clinical decision support systems could reduce the cost of health care delivery in my area of expertise.	1	2	3	4

16. I like having a percentage associated with each prediction: ☐ yes ☐ no.

17. Would information about the likelihood of long term disability be a desirable feature to include in future versions of the tool? ☐ yes ☐ no.

18. Would outcomes other than those in the current application be useful for decision making? If so, which ones?

19. Would a case-matching feature be desirable in a decision support system for use in the NICU? (E.g., Making the 10 most similar cases (data and outcomes) to the one under consideration available for viewing in a separate screen).

☐ yes ☐ no.

20. Other comments about the ANN tool.
(E.g., your opinion, what are the most positive benefits of the system? Do you see the possibility of anyone misusing the system? **Do you believe such systems could be beneficial to the delivery of care in the NICU?** Issues you feel were overlooked or not addressed in the best way.)

(Please feel free to attach further pages as necessary to capture your comments)

APPENDIX E – Informed Consent for Physicians

Division of Neonatology/Service de néonatologie
(613) 737-2393, extension/poste 2415
Fax/Téléc: (613) 738-4847

Comparison of Prediction by an Artificial Neural Networks Tool and Physicians of Outcomes of Neonatal Intensive Care Unit Cases

Explanation of the Study and Health Professional Consent

Erika Bariciak, University of Ottawa, Ottawa, ON 613-737-7600 ext 2415
Monique Frize, P Eng, OC, Carleton University, Ottawa, ON (613) 520-2600 X8229
Daphné Townsend, B A.Sc, University of Ottawa, Ottawa, ON (613) 562-5800 X6013

SPONSOR: None at this time
You are being asked to take part in a research study. Before agreeing to participate in this study, it is important that you read and understand the following explanation of the proposed study procedures. The following information describes the purpose, procedures, benefits, discomforts and risks associated with this study. It also describes your right to refuse to participate or withdraw from the study at any time. In order to decide whether you wish to participate in this research study, you should understand enough about its risks and benefits to be able to make an informed decision. This is known as the informed consent process. Please ask the study doctor or study staff to explain any words you don't understand before signing this consent form. Make sure all your questions have been answered to your satisfaction before signing this document.

Purpose
The overall goal of this research project is to compare the predictions of a proposed decision-support system for potential future clinical application using intelligent tools (e.g., artificial neural networks) with predictions by experienced neonatal physicians. The case data and prediction interface will be validated by means of a questionnaire to obtain physicians' attitudes and comments.

Procedures
You will be asked to predict the outcomes of a series of cases using data that has also been supplied to the 'intelligent' system. You will not receive any information that could allow you to identify real cases, although the data was derived from actual cases contained in a database used in our research. Your predictions and those of the system will then be compared with the actual outcomes. There will be no attribution of predictions to individual physicians nor will physician accuracy be graded or revealed in any way – the intent is simply to discover whether the new systems can offer additional

accuracy over predictions made by neonatal physicians in general or not. This will help assess the likely value, if any, of such new systems.

Prediction results will be analyzed as a group and will not be attributed to particular physicians or cases.

Benefits

There are no immediate benefits for your participation in this study. We hope that the results of this study will help us to develop and refine methods for use in practice and ultimately improve decision-making in the NICU.

Discomforts, Risks

There are no anticipated discomforts or risks with this study.

Confidentiality

All information in this study is voluntary. You can choose not to participate or you may withdraw at any time.

Questions

If you have any questions about the study, please call **Dr Erika Bariciak at 737-7600 ext. 2415**. If you have any questions about your rights as a research subject, please call Dr Carole Gentile, Chair of the CHEO Research Ethics Board at 737-7600 ext. 3624.

Consent

I have had the opportunity to discuss this study and my questions have been answered to my satisfaction. I consent to take part in the study with the understanding that I may withdraw at any time. I have received and signed copy of this consent form. I voluntarily consent to participate in this study.

_____	_____	_____
Participant's Name (Please Print)	Participant's signature	Date

I confirm that I have explained the nature and purpose of the study to the subject named above. I have answered all the questions.

_____	_____	_____
Name of Person Obtaining Consent	Signature	Date