INFORMATION TO USERS

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

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

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

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

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

UMI[®]

EVALUATION OF AN ARTIFICIAL NEURAL NETWORK TOOL FOR NEONATAL INTENSIVE CARE UNITS

by

Lili Qi

Supervised by Professor Monique Frize and Dr. Robin Walker

A thesis submitted to

The Faculty of Graduate Studies and Research

In partial fulfillment of the requirements for the degree of

Master of Science

In Information and Systems Science

Department of Systems and Computer Engineering

Carleton University

Ottawa, Ontario, Canada

May 2005

Library and
Archives Canada

Bibliothèque et
Archives Canada

Published Heritage
Branch

Direction du
Patrimoine de l'édition

0-494-08361-1

395 Wellington Street
Ottawa ON K1A 0N4
Canada

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file *Votre référence*

ISBN:

Our file *Notre référence*

ISBN:

NOTICE:

The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

AVIS:

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protègent cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.

Canada

Abstract

The Artificial Neural Network (ANN) prototype was developed using medical databases available at the CHEO (Children's Hospital of East Ontario) to help neonatologists in their decision-making. Nine SNAPPE-II (The Score for Neonatal Acute Physiology-Perinatal Extension-II) variables were used as the input variables of the ANN tool. Weights of the nine variables were trained using the CNN (Canadian Neonatal Networks) database. The system was designed to predict the likelihood of infants' survival, the duration of artificial ventilation, and length of stay in the Neonatal Intensive Care Unit (NICU).

Two pilot studies were conducted for its evaluation. One was an attitude test, which aimed to measure the physicians' attitudes toward using the ANN tool and collect opinions and suggestions for its improvement. The results of the study found that physicians' responses to the ANN tool were fairly positive. One variable, gestational age, which was currently not used by the tool, was identified as important in guiding physicians in predicting outcomes. The other pilot study was laboratory experiments, which aimed to evaluate the performance of the tool. The results of the experiments demonstrate that the best testing results of the ANN training using CNN database are better than the performance of the physicians on EPIC database in terms of sensitivity, the specificity and the CCR. When used with the EPIC database – in which the population was different from the CNN database – the ANN tool also showed a comparable performance to that of the physicians. This serves as a further indication that the ANN tool has the strength to help physicians with their decision-making.

In loving memory of my mom, Shufen Dong,
who supported my graduate study during her illness,
and
to my Dad, Nianzu Qi,
and my brother Jian Qi,
for their endless support and encouragement,
also
to my husband, Alexander Duan,
for helping me to overcome many difficulties.

ACKNOWLEDGEMENTS

I would like to express my deep appreciation to my supervisor Dr. Monique Frize, for guiding me throughout my graduate studies at Carleton University. Thank you for giving me intellectual instructions, spiritual encouragement and kind help which allowed me to complete my thesis for a Master's Degree.

My sincerest gratitude goes to Dr. Robin Walker for his frequent advice and help. I appreciate very much the time you provided for this research.

Many thanks to the neonatologists in the Neonatal Intensive Care Unit at Children's Hospital of Eastern Ontario and at the Ottawa Hospital-General Campus for their participation in the pilot study.

I appreciate very much the selfless, kind help provided by Doaa Swailum on neural network techniques. Thanks also to Dawn Rybchynski for her help on the development of the prototype of the ANN tool.

And most importantly, I would like to acknowledge the continuous support of my husband, Alexander Duan. A special thank for giving courage and help during the difficult times.

Contents

Chapter 1	Introduction.....	1
1.1	Motivation.....	1
1.1.1	The Neonatal Intensive Care Unit Environment.....	1
1.1.2	Medical Expert System.....	1
1.1.3	The application of Artificial Neural Networks in Intensive Care Units .	2
1.2	Research Status and Objectives	3
1.3	Summary of the Contributions.....	4
1.4	Thesis Outline	4
Chapter 2	Literature Review	5
2.1	Evaluation Methods for Medical Expert Systems	5
2.2	Theories and Structures of Artificial Neural Networks	10
2.2.1	Transfer Functions	15
2.2.2	Weight-elimination and Weight-decay	16
2.2.3	Adaptive Network Parameters	16
2.2.4	Stopping Criteria.....	18
2.2.5	Performance Measurement	19
2.2.5.1	Constant Predictor.....	19
2.2.5.2	Correct Classification Rate	20
2.2.5.3	Sensitivity and Specificity	21
2.2.5.4	Receiver Operating Characteristic (ROC) Curve	21
2.3	Application of Artificial Neural Networks	23
Chapter 3	Introduction of ANN tool	25
3.1	Score for Neonatal Acute Physiology.....	25
3.2	SNAP-II and SNAPPE-II.....	27
3.3	About Canadian Neonatal Network (CNN) Database	28
3.3.1	The CNN Database and Missing Values in CNN Database	28
3.3.2	Imputation of Missing Values by Integrating ANN and CBR.....	30
3.4	ANN Tool Trained with 5088 Cases from CNN Database.....	31

3.5	Training ANN Tool with 19398 Cases from CNN Database	32
3.5.1	Re-training the ANN Tool	32
3.5.1.1	ANN Architecture	32
3.5.1.2	Preparation of Data	33
3.5.1.3	ANN Implementation Process	34
3.5.2	ANN Training Results	36
3.5.3	ANN Tool	39
Chapter 4	Methodology	42
4.1	Pilot Study 1–Attitude Test.....	43
4.1.1	Design of the Attitude Test.....	43
4.1.2	Implementation Process of the Attitude Test.....	45
4.1.2.1	Subjects	45
4.1.2.2	Procedure	45
4.2	Pilot Study 2: Laboratory Experiment	46
4.2.1	Design of the Experiment	46
4.2.2	Implementation Process of the Experiment	46
4.2.2.1	Data Requirements:.....	46
4.2.2.2	Data Sources:	48
4.2.2.3	Extracting Data from Database and Data Preparation:	48
4.2.2.4	Predictions of Physicians	52
4.2.2.5	Predictions of ANN Tool	52
Chapter 5	Results and Discussion	53
5.1	Results and Analysis for the Attitude Test	53
5.2	Results and Analysis for the Lab Experiment.....	62
5.2.1	Results and Analysis for the ANN Tool	62
5.2.2	Results and Analysis for Physicians’ Predictions.....	69
5.3	Discussion	72
5.3.1	Strengths and Limitations of the ANN Tool in Making Predictions	73
5.3.2	Strengths and Limitations of Physicians in Making Predictions	75
Chapter 6	Conclusions, Contributions and Future Work.....	76
6.1	Conclusions.....	76
6.2	Contributions.....	77

6.3	Future Work.....	77
	References.....	79
Appendix 1	SNAP Variable List and Scoring System	86
Appendix 2	Questionnaire for Pilot Study 1	87
Appendix 3	The Specifications for SNAP	90
Appendix 4	Proposal for Project	91
Appendix 5	Consent Form for Participating the Project	94

List of Tables

Table 2- 1 Pseudocode for neural network parameters.....	18
Table 2- 2 Contingency table.....	20
Table 3- 1 Definitions of SNAP point assignment [Richardson et al. 1993].....	26
Table 3- 2 SNAP-II and SNAPPE-II variable list and scoring system [Richardson <i>et al.</i> 2001]	28
Table 3- 3 Classification Performance of ANNs: Mortality, Ventilation and Length of Stay in the NICU [Frize M. 2003].....	32
Table 3- 4 Proportion of cases in the CNN database	33
Table 3- 5 Classification performance for Mort and LOS	36
Table 3- 6 Classification performance for Ventday.....	38
Table 3- 7 Weights and biases used in the ANN tool.....	39
Table 4- 1 Patient's information needed for physicians	47
Table 5- 1 Responses to the part 1 of the questionnaire	53
Table 5- 2 Rank of variables by each physician	57
Table 5- 3 Variables chosen by physicians and the rank.....	58
Table 5- 4 Answers for question 10.....	59
Table 5- 5 Answers for question 11.....	61
Table 5- 6 Correct prediction case # of ANN tool for EPIC	62
Table 5- 7 Classification performance of ANN tool (best testing results from training ANN using CNN database) & its prediction rates for EPIC database.....	62
Table 5- 8 Case proportions of EPIC database	63
Table 5- 9 statistical values from EPIC, test and training datasets.....	64
Table 5- 10 Probability of t-Test for EPIC, Test and Training Sets	64
Table 5- 11 ANN tool variable weights.....	65
Table 5- 12 Test results for variable "Mort".....	66
Table 5- 13 Test results for variable "LOS "	66
Table 5- 14 Test results for variable "Ventday" (2-layer)	66
Table 5- 15 Test results for variable " Ventday" (3-layer)	67

Table 5- 16 Classification rate of ANN tool (best results from testing ANN using CNN database) & classification rate for EPIC by physicians.....	69
Table 5- 17 Real outcomes for 59 cases and cases correctly predicted by physicians and by ANN Tool.....	70
Table 5- 18 Comparison of physicians' and tool's prediction results for each	72

List of Figures

Figure 2- 1 A Basic Artificial Neuron [Anderson D. <i>et al.</i> 1992]	11
Figure 2- 2 Two-layer ANN	12
Figure 2- 3 Three-layer ANN	13
Figure 2- 4 Sample ROC Curves [ROC onlin1]	22
Figure 3- 1 ROC curve for “Mort”	37
Figure 3- 2 ROC curve for “Los”	37
Figure 3- 3 ROC curve for “Ventday” (3-layer).....	38
Figure 3- 4 ROC curve for “Ventday” (2-layer).....	39
Figure 3- 5 Interface of the ANN tool	40

Abbreviations

ANN	artificial neural network
APACHE	acute physiology and chronic health evaluation
ASE	average squared error
CBR	case-based reasoner
CCR	correct classification rate
CHEO	children's hospital of eastern Ontario
CNN	Canadian neonatal network
CP	constant predictor
EPIC	evidence-based practice identification and change
FN	false negative
FP	false positive
ICU	intensive care unite
LOS	length of stay in hospital
MIRG	medical Information technology research group
NICU	neonatal intensive care unite
PSI	physiologic stability index
ROC	receiver operating characteristic
SNAP	score for neonatal acute physiology
SNAP-II	score for neonatal acute physiology generation two
SNAPPE-II	score for neonatal acute physiology perinatal extension generation two
SSE	sum of squared error
TN	true negative
TP	true positive
TSSE	training set SSE from the current approach

Chapter 1 Introduction

1.1 Motivation

1.1.1 The Neonatal Intensive Care Unit Environment

Neonatal intensive care units (NICU) are designed with special equipment to care for premature, low-birth-weight and/or seriously ill newborns. Neonatal medicine has been available in developed countries for more than three decades to provide specialized and intensive care measures aimed at improving the health and survival of premature and critically ill newborns [Carter BS, 2004]. Throughout these years, many advances have been made, such as the substantial reduction in the mortality of premature infants, and the steadily declining rate of disability or significant morbidity in survivors of the NICU of nearly all gestational ages and weights [Carter BS, 2004]. However, according to the author, the rate of prematurity, low birth weight and the rate of birth defects have remained relatively constant over the past few decades.

1.1.2 Medical Expert System

The computer-aided expert system can extract, formalize and store much information provided by human experts, which in turn, can apply the acquired knowledge in a particular problem-solving task by emulating the reasoning process of human experts. Such computer-aided expert systems in medicine are expected to provide physicians with useful information from large numbers of patient records. A well designed expert system can improve the quality of patient care by helping clinicians in making a diagnosis and treatment plan. It can also help physicians to allocate resources and reduce the overall

costs of health care. In recent years, due to the explosion of technologies, especially in computer science and engineering, scientists and engineers have shown interest in medical expert systems, neural networks and other artificial intelligence techniques. Healthcare providers also realize their potential. Medicine and healthcare are a popular field for developing new or applying existing technologies.

1.1.3 The application of Artificial Neural Networks in Intensive Care Units

The application of Artificial Neural Networks (ANNs) in Intensive Care Units (ICU) includes evaluating clinical parameters, estimating medical outcomes, resource utilization and management of treatment. ANN can support physicians by initiating necessary responses in case of emergency (within 17 seconds instead of the 45-second average by clinicians) [Kornel P. *et al.* 1994]. [Buchman *et al.* 1994] reported the use of a ANN to evaluate the relationship of several demographic, pharmacologic, and physiologic variables with chronicity in a surgical ICU. A ANN model with fifteen input variables developed by Tu and Guerriere [1993] is reported to estimate the length of stay and mortality in cardiovascular ICUs. Snowden *et al.* [1993; 1997] developed an ANN system with two hidden layers, each composed of fifty nodes to manage ventilated neonates in the NICU. The system takes blood-gas parameters and the current ventilator settings as inputs and outputs a set of suggested new ventilator settings.

Frize *et al.* began research on ANN for intensive care in the early 90s. A feed-forward ANN was trained using the back-propagation algorithm to estimate mortality, duration of ventilation and length of stay in the adult ICU [Buskard T. *et al.* 1994]. In 1995, the MIRG (Medical Information technology Research Group) [Frize M. *et al.* 1995]

developed the application of intelligent monitoring techniques of case-based reasoning and ANN analysis to physician decision making concerning patient care in an ICU. Walker *et al.* [2000] applied a three-layer feed-forward back-propagation ANN to predict ventilation duration less than or, more than or equal to 8, 12, and 24 hours and death on a large database of NICU patients. An evidence-based ethical decision-making tool [Frize M. *et al.* 2004; Lang Y. *et al.* 2004]], combined with two artificial intelligent approaches, case-based reasoning and ANN, was developed in early 2005 [Frize *et al.* 2005]. The system was designed to provide more information to clinicians and parents when they are facing very difficult, ethical decisions regarding the care or management of neonatal intensive care patients.

1.2 Research Status and Objectives

The long term objective of the MIRG is to use artificial intelligence approaches to create a number of promising clinical decision-support systems which will ultimately be integrated into one main system. With the efforts of many research projects, the ANN and CBR systems have been developed to a point of clinical applicability.

The objective of this thesis was to evaluate MIRG's ANN tool for its usefulness in neonatal intensive care unit (NICU), to see whether the tool was able to help physicians in estimating outcomes and counseling parents. This was a pilot study which was employed as a preparation for a large scale clinical trial in the future.

1.3 Summary of the Contributions

The objective of the thesis was met through the implementation of an attitude test to measure the physicians' attitude on using the ANN tool, and a laboratory test to evaluate the performance of the ANN tool. Since this study involved some ethical issues, a proposal of this project has been developed and submitted to the Ethics Review Committee at CHEO (Children's Hospital of Eastern Ontario), and was subsequently approved. The results of the study show that the ANN tool is fairly welcomed by the physicians, and the performance of the ANN tool is comparable with that of the physicians. From the results, a key variable was identified (gestational age), which is important for physicians to do predictions, but was not included in the input variables of the ANN tool. Also, several strengths and limitations of the ANN tool were identified from analysis of the results, which can form the basis for future work.

1.4 Thesis Outline

The thesis consists of six chapters. Chapter1 outlines the motivation, research status and the objective. A literature review of related subjects to the thesis is given in chapter 2. It includes evaluation methodology for medical expert system and overview of ANN. Chapter3 provides a detailed introduction of ANN tool that needs to be evaluated. The methodology for the evaluation is addressed in chapter 4. There are two parts: one is from the user's perspective, measuring the attitudes and collecting user's opinion for the tool's improvement; the other focuses on the performance of the ANN tool by comparing the predictions made by physicians and by ANN tool, as well as real outcomes. Chapter 5 summarizes the results of the study and presents an analysis and discussion for the

results. In the last chapter, conclusions of the study are presented. Contributions to new knowledge and suggestions for future work are also supplied.

Chapter 2 Literature Review

This chapter reviews literature on evaluation methods of medical expert systems which led to the selection of a method for the evaluation of our ANN tool. Also, theories, structures, and applications of ANNs are briefly introduced to provide a general idea of ANNs, and to help to understand the training process of ANN discussed in the next chapter.

2.1 Evaluation Methods for Medical Expert Systems

Today, health care depends more and more on advanced technologies, especially on computer systems. However, there are risks that the use of computer expert systems in medicine could reduce the quality of health care. It is clear that usually there are bugs in soft wares. Excessive reliance on systems, overestimation of their performance, and insufficient evaluation and maintenance are factors causing misdiagnosis, mistreatment and even fatalities [Leveson N., 1993]. For these reasons we first discuss the importance of evaluating medical expert systems before the overview of evaluation methods.

Three major reasons — ethical, legal and intellectual — have been mentioned by Wyatt *et al.* [1990] in discussing the necessity of performing evaluations. The ethical basis of medicine refers to the improvement of patients' health while attempting to do no harm and using limited health care resources wisely. Therefore, the evaluation should focus on whether an expert system is effective and safe and how it might change usage of

resources, before attempting to distribute it. Expert systems technology is also subject to litigation. “Product liability law dictates that it must be safe, no matter what precautions have been taken to avoid mishaps” [Wyatt J. *et al.* 1990]. The intellectual arguments for testing expert systems include: which technical advances lead to progress; which domains are the most fruitful to pursue; and how to predict the performance of a knowledge-based system from its structure.

Reisman [1996] points out additional purposes for which the evaluation should be performed: to provide a concrete test of a prototype system; to obtain feedback for further system development and refinement; and to determine the system’s clinical effects, that is, whether the system indeed achieves its goal; and to develop new evaluation methodologies.

In the overview on the evaluation methodology of knowledge-based systems, [Wyatt J. *et al.* 1990] summarize the evaluation of medical expert systems in three stages. “The first stage consists solely of a definition phase, which identifies the exact decision problem, the users and their environment. To achieve a complete definition of user requirements, an early prototype system should be built, and users should be requested to comment on how this could be improved. Thus, stage one may prove to be an iterative build-test-refine cycle.”

In the second stage, further prototype systems are built. At this stage, formal laboratory testing should begin before the system can be introduced into clinical practice. “This

laboratory testing phase is necessary to show that the system is safe and has at least the potential to benefit patients. The system at this time is viewed as a 'decision taker', in that its conclusions are judged directly. Only in the next phase of evaluation will its effects as an aid to clinical decisions be considered." [Wyatt J. *et al.* 1990]

The structure, process and outcome of the system should be tested. Questions about knowledge-based structure and reasoning need to be answered by those with insight into both the technology and the domain [Pearl J. 1988].

"From the user's perspective, we must have a means of measuring attitudes and 'user satisfaction' and this may be difficult to do formally because of their subjectivity. It can be done with a structured questionnaire combining verbal scales with visual analogue scales." [Reisman Y. 1996] Questions that could be asked are the following:

"Structure (Is the system wanted?)

Is there a perceived need?

Can the advice be made available at the right place and time?

Process (Is it pleasant to use?)

Is the interface satisfactory?

Is the system fast enough?

Outcome (Does it say sensible things?)

Do its conclusions seem as sound as those of a respected authority?

Are the explanations adequate?" [Wyatt J. *et al.* 1990]

To measure the accuracy of a system an appropriate test sample is required. Data used for debugging a system should not be used for laboratory testing since any expert system will usually perform much better on such 'training' data than on a fresh set of data, collected as part of a separate study [Wasson JH. *et al.* 1985].

The third stage of evaluation consists of field trials, which are crucial to determining if the expert systems have value for patients or for decision-takers [Wyatt J. *et al.* 1990]. They can be designed to check whether the intervention has an effect on outcome (a pragmatic motive) or to elucidate how it exerts this effect (explanatory motive) [Schwartz D. *et al.* 1967]. "It seems reasonable to examine the effect of a system on the structure (how well does the system fit into its intended environment), on the process (what effects does the system have on the process of health care) and on the outcome (does use of the system have an effect on outcome measures) of the health care process, controlling for the effects of bias and of confounding variables." [Johnston ME. *et al.* 1994; Nohr C. 1994]

Clarke *et al.* [1994] place emphasis on the importance of safety, correctness and the users' confidence in the knowledge-based expert systems. They describe an iterative, four-phased development evaluation cycle covering the following areas: early prototype development, evaluation of validity of the system, functionality of the system, and field trials where the impact of the system is evaluated. In phase two, the evaluation of the system's validity needs to be determined during the implementation of the system. Validity assessment focuses on the correctness of the output of the system. The authors

point out several methods for assessing outcomes: the traditional error rate approach is a simple method whereby the rate of correct answers of the system is measured. This does not take into account systems that give answers in terms of probability, or those that rank their answers in order of likelihood; top-down correlation compares the ranked preferred solutions of the expert(s) with those of the system. It has been shown to be an excellent form of correlation as it places more emphasis on the top of an expert's rank-ordered list than on the bottom; ROC (Receiver Operated Characteristics) is a plot showing the covariation of the sensitivity and 1-specificity as a function of the discriminator value. The ROC function is independent of the prevalence of the actual disorder and has been used in the evaluation of knowledge based systems [Miller RA. *et al.* 1982]. The optimum of these plots is a compromise between false positive and false negative findings — the compromise being dictated by the application of the system.

Adelman *et al.* [1997] Riedel give five types of evaluations including requirements validation, knowledge-base validation, knowledge-base verification, usability evaluation and performance evaluation. Requirements validation and performance evaluation address the same very basic evaluation question: “will the operational system meet organizational and user goals and needs?” [Adelman L. *et al.* 1997] Both types of evaluation are predicting to what extent performance of the final system will match what is needed. The requirements are a stand-in for the actual system until development has proceeded to the point where the developing system can be assessed. Performance evaluation is more specifically concerned with whether performance using the system is more timely or results in an improved quality of the task process and products. The

quality of the knowledge base is assessed in two types of evaluations: knowledge-base verification and knowledge-base validation. "Knowledge-base verification looks at the logical consistency and adequacy of the knowledge base. Knowledge-base validation, on the other hand, is the discovery of potential errors in the logical content and structure of the knowledge base." [Adelman L. *et al.* 1997] Usability evaluation helps determine "how easy a system is to learn, the number of errors users are likely to make in using the system, whether or not users will have confidence in the system results, the amount of cognitive effort system operation takes, and even whether users will use the system." [Adelman L. *et al.* 1997] Knowledge-base verification and validation, as well as usability evaluation are more concerned with identifying problems that might interfere with the optimal performance of the system. Each of the five types of evaluation can be conducted using many different methods. Unfortunately, the evaluator usually does not have his or her choice of all of these methods because the choice of methods and sources of data are constrained by a number of different factors, such as "the amount of evaluation funding, availability of SMEs (subject matter experts) and users, time to plan the evaluations, number and expertise of evaluation personnel, cooperation of sponsors and developers, and opportunity for the evaluator to learn about users and organizational units using the system." [Adelman L. *et al.* 1997]

2.2 Theories and Structures of Artificial Neural Networks

In medical areas, techniques of ANNs have been widely used. The main advantage of ANN over statistical techniques is that the model does not have to be explicitly defined before beginning the experiments to develop the model. ANN have a learning ability and can perform successfully where statistical techniques do not, in recognizing and matching

complicated, vague, or incomplete patterns [Blum 1992, Livingstone *et al.* 1997]. Also, with statistics, it is difficult to integrate data of different formats (i.e., working simultaneously with continuous, binary, ordinal and nominal data). Medical data are often believed to have nonlinear relationships between the input variables and the outcome, so multilayer ANN are a logical choice for analysis in the medical environment [Baxt 1994].

An ANN is a parallel, distributed, adaptive system that processes information in order to learn patterns as a result of exposure to a set of representative training cases [Penny W. *et*

Figure 2- 1 A Basic Artificial Neuron [Anderson D. *et al.* 1992]

al. 1996]. The original intention was to use ANNs to emulate human brain function through the structure of nodes. In the biological sense, a neuron is “a cell with long processes specialized to receive, conduct, and transmit signals in the nervous system”

[Alberts B. *et al.* 1998]. There are about 100 billion neurons in the human brain. Learning is accomplished by the adjustment of the synaptic connections between adjacent neurons. ANNs learn the solution to problems by adjusting the connecting weights to output appropriate results.

The artificial neuron (Fig. 2-1) is a simplified mathematical model of biological neurons. It is the basic unit of ANNs. Each artificial neuron, which is called a node in ANNs, has many inputs ($x_0 - x_n$), a processing unit and one output (y). The processing unit sums the products of inputs and their corresponding weights ($w_0 - w_n$), and the signal passes through a transfer function to yield a result (y).

In ANNs, nodes are grouped into three kinds of layers: input layer, hidden layer(s) and output layer. An ANN can have two layers (input and output layer) or more (hidden layer(s)) as shown in Fig. 2-2 and Fig. 2-3. Nodes in the input layer receive input

Figure 2- 2 Two-layer ANN

Figure 2- 3 Three-layer ANN

information from the external environment and yield input signals to the other layer(s). The outputs of the system can be obtained from the nodes in the output layer. The signal processing path is normally unidirectional, but there may be a two-way connection between layers. The different types of connections between layers create various ANNs, which include fully-connected, partially-connected ANNs. In more complex structures, intra layer connections can be observed. The nodes within a layer are fully or partially connected to one another. A feed-forward network allows information to travel directly from the inputs to the outputs. If any outputs are fed back as inputs to previous layers, this is a feedback or a recurrent network [Fausett L. *et al.* 1994].

Hidden layers are introduced to solve more complicated problems using non-linear algorithms; however, training time will significantly increase with the increase of hidden layers. The good news is that not all networks require a hidden layer, and no substantial

evidence has been found showing that more than one hidden layer will improve the predictive performance of a network [Penny W. *et al.* 1996].

Caution must be taken when choosing the number of nodes in the hidden layer(s). This is a trade-off situation. In order to model the dataset closely, more nodes should be added to the network. However, a model that is too closely related to the training set is unable to generalize to accommodate new data [Fausett L. *et al.* 1994]. Also it may reduce the efficiency of the system since having too many nodes in the hidden layers would use considerable computing resources. On the other hand, too few nodes will not provide sufficient resources to solve the problem, resulting in underfitting. Underfitting means that the ANN does not learn enough about the dataset to correctly classify cases.

Like the human brain, ANNs can learn from experience as more data are added into the training dataset. Applying the new training data to the chosen algorithm will generate new connection weights in the network. In this way, the system learns new knowledge by adjusting these connection weights. The learning ability of ANN is determined by its architecture and by the algorithmic method selected for training. The training can generally be grouped into unsupervised and supervised learning. Unsupervised training means that the true output is unknown and the nodes must find a way to organize themselves without help from the outside. In this approach, no sample outputs are provided to the network against which it can measure its predictive performance for a given vector of inputs. This is so-called learning by doing. For supervised training, the true output is known and the ANN uses this information to reshuffle the connections and

correct its errors. This reinforcement learning is also called learning by example.

[Anderson D. *et al.* 1992]

There are a variety of learning rules in common use. These rules are mathematical algorithms used to update the connection weights. The feedforward, error back-propagation algorithm, also known as gradient descent learning, was first introduced by Rumelhart *et al.* in 1986, and since then a number of computational enhancements have been developed to improve the algorithms [Mehrotra K. *et al.* 1997]. This method is among the most commonly used network training algorithm [Werbos P. *et al.* 1974]. It uses the error between the target (true output) and system output to modify the weights of the network. The error is propagated back through the layers, adjusting the node weights to minimize the overall error or to reinforce the important weight [Livingstone DJ. *et al.* 1997].

2.2.1 Transfer Functions

There are several transfer functions which are commonly used, such as the hard-limit, the linear and the log-sigmoid transfer function. The log-sigmoid transfer function is often used in the back-propagation networks. The hyperbolic tangent, or tan-sigmoid function, differs from the former in its sharp transition between its output values that permits faster learning of ANNs. It is also a common transfer function in ANNs for the reasons that it is a very flexible, non-linear function, and is also continuous and differentiable.

2.2.2 Weight-elimination and Weight-decay

Weight-elimination and weight-decay add a complexity term to the error function to prevent overfitting to the training set [Weigend AS. *et al.* 1990]. Overfitting means the networks are too specific to the training data to extend to the prediction of general cases. The weight decay penalty term causes the weights to converge to smaller absolute values than they otherwise would. Large weights can cause excessive variance of the output [Geman S. *et al.* 1992]. Therefore, weight decay results in a more stable network. According to Bartlett [1997], the size of the weights is more important than the number of weights in determining generalization. On the other hand, weight-elimination focuses on the small weights by trying to reduce them to zero. The small weights only add unwanted “white noise” to the model. This network pruning approach simplifies the network and can eliminate variables that offer little or no assistance in predicting the correct outcome [Weigend AS. *et al.* 1990].

2.2.3 Adaptive Network Parameters

Network parameters need to be adjusted to gradually reach the optimal performance of the networks and to develop the best ANN model. The error ratio *err_ratio* is a user defined value that measures the error change from current to previous for the determination on how the learning rate *lr*, momentum *m* and the weight decay constant λ (or *lambda*) are updated. The objective is to have the error ratio get smaller with each iteration to show that the performance is improving [Fausett L. *et al.* 1994]. The learning rate controls the size of steps across the decision surface. By adjusting it during the training process through the parameter *lr_inc* and *lr_dec*, the optimal performance of the network can be reached.

Table 2-1 shows the pseudocode for how parameters are adjusted [Fausett L. *et al.* 1994].

TSSE refers to the training set error from the current epoch (an epoch is defined as a single pass through the entire training set, followed by testing of the verification (testing set); SSE is the error from the previous epoch. The use of momentum is to prevent the

```
if TSSE>SSE*err_ratio
    lr=lr*lr_dec;
    lambda=lambda*lambda_dec;
    m=0;
else if TSSE<SSE*err_ratio
    lr=lr*lr_inc;
    lambda=lambda*lambda_inc;
end
```

Table 2- 2 Pseudocode for neural network parameters [Fausett L. *et al.* 1994]

training from getting caught in local minima and allows it to continue to search for the global minimum. Just like a low-pass filter disregards low-level noise, momentum follows the general flow of the error surface by ignoring any small feature [Fausett L. *et al.* 1994]. It is implemented by adding a fraction of the last weight change to the new weight change. If the current-to-previous iteration error is greater than the error ratio, then the momentum is set to zero and the new weights and biases are rejected. Otherwise the new weights and biases are retained and the weight change is calculated [Fausett L. *et al.* 1994].

The role of the weight decay constant (λ) is to determine how strongly weights are penalized. Weight elimination adds the following penalty term in addition to the main mean squared error function [Weigend AS. *et al.* 1990]:

$$\lambda \sum (w_i/w_0)^2 / [(w_i/w_0)^2 + 1]$$

where w_i represents the individual weights of the ANN and w_0 is the weight-decay scaling parameter. The value of λ is user-defined. The larger the value of λ , the closer the weights must be to zero to be considered among the “noise” distribution. A larger λ also puts more pressure on smaller weights to go to zero [Fausett L. *et al.* 1994]. If the value is too small, there will be no impact on the network; however, if it is too large, all weights will be forced to zero [Fausett L. *et al.* 1994]. Weigend *et al.* [1990] suggested initially setting the value of λ to zero to allow the network to gather information for all input variables, then gradually increasing its value until the performance begins to degrade, at which point we fine tune the λ value to optimize the performance.

2.2.4 Stopping Criteria

The automated network [Frize M. *et al.* 2000, Ennett CM *et al.* 2004] was developed to optimize each network parameter without user supervision. A stopping criterion determines when the automated network has reached a satisfactory level of performance and needs to stop training. Several stopping criteria are used by MIRG, such as correct classification rate (CCR), Average Squared Error (ASE) and the log-sensitivity index [Scales 2001, Ennett CM *et al.* 2002]. Also, an early stopping criterion is employed in the training program combined with each of the above stopping criteria. It stops the training when the performance does not improve during the subsequent 500 epochs in order to prevent overfitting. Using the CCR as stopping criterion, the network stops iterating after reaching the highest correct classification rate, while ASE lets the network stop when the lowest average square error is found. Both can be good measures of performance when

outcomes are relatively common (representatively of approximately 15 percent or more of the database). [Ennett CM. *et al.* 1998]

The logarithmic-sensitivity index (or log-sensitivity index) is introduced as a stopping criterion when working with data where the outcome is a rare occurrence (represented by only a few percent of the cases in the database). It is defined as the following:

$$\text{log-sensitivity index} = -\text{sensitivity}^n * \log_{10}(1 - \text{sensitivity} * \text{specificity})$$

This function expands towards infinity as sensitivity and specificity approach one, but is slightly weighted in favour of sensitivity [Fausett L. *et al.* 1994]. The degree of weighting can be controlled by the exponent n [Scales N. 2001]. Studies have shown that an automated ANN which optimizes the log-sensitivity index can classify outcomes better than manually-optimized networks, and can classify rare outcomes equally well or better than networks that use the maximum correct classification rate or the minimum mean squared error as the stopping criteria [Ennett CM. *et al.* 2002; 2003].

2.2.5 Performance Measurement

To evaluate the performance of networks needs appropriate measurements [McGowan HCE, 1996]. There are several performance measurements for ANN which are commonly accepted and used.

2.2.5.1 Constant Predictor

A constant predictor is a basic statistical benchmark that classifies all cases as belonging to the class with the highest *a priori* probability [Frize M. *et al.* 1995]. To illustrate this, consider the Canadian Neonatal Network (CNN) database. The percentage of deaths is

3.7, while the survival is 96.3 percent. With this data set, a constant predictor would classify all patients as surviving, and it would correctly classify the patients 96.3 percent of the time.

2.2.5.2 Correct Classification Rate

Correct classification rate (CCR) identifies the rate at which the model correctly classifies the data into their proper categories. Based on the contingency table below, CCR or accuracy, is calculated by sum of true positive (TP) and true negative (TN) divided by the total number of cases in the dataset. The calculating formula is $(TP+TN)/(TP+TN+FP+FN)$. TN and TP represent cases that are correctly classified into their respective classes, while false positive FP and false negative FN represent those incorrectly classified cases. Though CCR is a good measure of network performance, it can not reflect the whole aspect of the classification especially when the data are highly skewed. Therefore, it is better to combine this with other measures such as CP, sensitivity, specificity and ROC curves to obtain a more valuable performance evaluation.

Table 2- 3 Contingency table

		Real output		
		Negative (N)	Positive (P)	
Predicted output	N	True negative(TN)	False negative(FN)	TN+FN
	P	False positive(FP)	True positive(TP)	FP+TP
		TN+FP	FN+TP	Total cases

2.2.5.3 Sensitivity and Specificity

The network's sensitivity is the proportion of positive cases that are correctly identified. Therefore, the formula for sensitivity is $TP/(TP+FN)$, where $TP+FN$ represents the total number of positive cases. Specificity is the proportion of negative cases that are correctly identified. The calculating formula is $TN/(TN+FP)$, where $TN+FP$ represents the total number of negative cases.

2.2.5.4 Receiver Operating Characteristic (ROC) Curve

The ROC curve shows the performance as a trade off between sensitivity and specificity. It is a plot of the model's sensitivity versus one minus its specificity (false positive rate). Constructing an ROC curve on a dataset allows one to identify how well a model discriminates between positive and negative results. The curve always goes through two points (0, 0) and (1, 1), where (0, 0) means no positives are detected and (1, 1) represents a classifier that classifies all cases as positives. The upper left hand corner, point (0, 1), represents a perfect classifier, which identifies correctly all the positive cases as positives and all the negative cases as negatives. Point (1, 0) represents the worst classifier, which cannot identify any case correctly. Therefore, a good ROC curve is the one that climbs rapidly towards the upper left hand corner of the graph, and far away from the 45-degree diagonal line which corresponds to a useless classifier. This means that true positive rate is high and the false positive rate is low.

Figure 2-4 [ROC online1] shows three ROC curves representing excellent, good, and worthless tests plotted on the same graph. Accuracy is measured by the area under the

ROC curve.

Figure 2-4 Sample ROC Curves [ROC online1]

An area of 1 represents a perfect test; an area of 0.5 represents a worthless test. A rough guide for classifying the accuracy of a diagnostic test is the traditional academic point system [ROC online2]:

0.90-1 = excellent (A)

0.80-0.90 = good (B)

0.70-0.80 = fair (C)

0.60-0.70 = poor (D)

0.50-0.60 = fail (F)

2.3 Application of Artificial Neural Networks

Not until the 1940s did the concept of the ANNs come into being, when W. McCulloch, a neurophysiologist, and a young mathematician, W. Pitts, published a paper on how neurons might work [McCulloch WS. *et al.* 1943]. With the invention of the computer, it became possible to simulate an ANN. Since the early nineties of the last century, ANNs have played an increasing role in the development of new biomedical systems, especially in the four basic categories: modeling, bio-signal processing, diagnosing, and predicting.

The application of ANNs has been introduced into almost every field of medicine. In 1990, Baxt [1990] used ANN as an aid to diagnose acute coronary occlusion and later for myocardial infarction [1991]. In order to classify heart failure patients, F. Atienza, *et al.* [2000] estimated three outcomes (death, readmission and one-year event-free survival) by applying ANNs. Although their sample size was limited, the total accuracy of predictions was over 93 percent.

Ravdin, *et al.* [1992] developed ANN system to judge the possible recurrence rate of tumors by using data from lymphatic node positive patients. The accuracy reached 96%. Fogel *et al.* [1995] reported their work on ANN recognition of breast cancer that evaluated mammographic, cytological and epidemiological findings in an integrated system.

In the neurological field, ANNs can be used to differentiate Alzheimer disease and vascular dementia. R. DeFigueiredo *et al.* [1995] analyzed brain SPECT image data with the help of ANNs. Kuntz *et al.* [1994] developed a cascade-correlation ANN model to

predict mortality and length of stay for patients with closed-head injuries. Walczak S. *et al.* [2001] evaluated the efficacy of using ANNs for determining epileptic seizure occurrences for patients with lateralized bursts of theta (LBT) EEGs.

Because of its ability to deal with large databases, ANNs have also been applied in the radiological field. By selecting an appropriate training set and learning process, ANN modeling becomes suitable for filtering, reconstruction and recognition of images, which contain so much information that it is impossible to interpret them using conventional rule-based systems. For the diagnosis of pulmonary embolism, Tourassi *et al.* [2000; 2001] presented an automated texture analysis method by calculating the complexity or “roughness” of each image. Based on ultrasonographic and laboratory findings, a differentiator [Maclin PS. *et al.* 1992] powered by ANN was created to diagnose five classes of liver disease.

There are also publications on applications of ANNs to the recognition of a particular pathology. To examine features extracted from light microscopic images, Dawson *et al.* [1991] developed ANN and established a classification process of breast carcinoma. To screen breast cancer and estimate its risk, analysis of DNA flow cytometric histograms are helpful. Nazeran *et al.* [1995] reported a study on increased breast cancer diagnostic precision by using ANN analysis.

Chapter 3 Introduction of ANN tool

The ANN tool was trained using the Score for Neonatal Acute Physiology Perinatal Extension-II (SNAPPE- II) as input variables. The training data were sampled from the Canadian Neonatal Network (CNN) database. The tool was trained with 5088 complete cases (without any missing values) by a graduate student in the MIRC. In this study, the tool was trained with 19398 cases in CNN database, where missing values had been replaced using MIRC's ANN and CBR (Case-Based Reasoner) hybrid system. In this chapter, more details about the ANN tool and its training processes we performed during this study are provided.

3.1 Score for Neonatal Acute Physiology

The Score for Neonatal Acute Physiology (SNAP) was developed as a neonatal illness severity index in the early 1990's. SNAP scores the "worst physiologic derangements in each organ system in the first 24 hours" [Richardson DK. *et al.* 1993]. It aims to make comparisons of differences in the substantial variation in birth-weight-adjusted mortality among newborns in NICUs. SNAP showed little correlation with birth weight and was highly predictive of neonatal mortality even within narrow birth weight strata.

The modeling technique was based on the scoring systems: APACHE (Acute Physiology and Chronic Health Evaluation) and PSI (Physiologic Stability Index), which had been developed for measuring the severity of illness in the adult or pediatric intensive cares. Variables selected for SNAP were modified from PSI for the purpose of adapting it to

neonates' particular characteristics. The specifications for SNAP [Richardson DK. *et al.* 1993] are provided in appendix 3.

Points of the SNAP scoring system were assigned to each variable according to deviation from the normal physiological value. The scale ranges from 0 to 5, with the definitions listed in Table 3-1. Missing data take a 0 value on the assumption that if such results were

Table 3- 1 Definitions of SNAP point assignment [Richardson et al. 1993]

Points	Definition
0	Normal physiologic range/Missing value
1	Physiologic deviation was sufficiently abnormal to merit careful monitoring
3	Deviation was such that most clinicians would alter therapy to correct it
5	Acutely life-threatening

important for clinical management they would have been measured and recorded [Richardson DK. *et al.* 1993]. Some values include both high and low measures for a particular physiologic variable in order to represent more extreme physiologic instability. SNAP consists of 28 variables (Appendix 1) of which 9 variables have both high and low values, resulting in 37 input variables being used to assess mortality risk. Scores ranged from 0 to 42, with a mean of 8.7. The more severe the condition, the higher the points scored. The score not only directly predicts in-hospital mortality, but is also highly correlated with several measures of medical resource use, including therapeutic intensity, nursing workload and LOS (length of hospital stay).

3.2 SNAP-II and SNAPPE-II

As a first-generation newborn illness severity score, SNAP was cumbersome to use because of the number and complexity of variables. Richardson et al. [Richardson DK. *et al.* 2001] modified the original SNAP score and created a second-generation SNAP score, which would have the following advantages:

1. **Simpler:** by reducing the number of score variables;
2. **More reliable:** by eliminating score variables that are complex or difficult to define and abstract;
3. **Empirically weighted:** by deriving the score empirically so that score variables reflect actual mortality risk rather than the original clinician estimates;
4. **Recalibrated:** by updating mortality risk estimates to reflect experience in the mid-1990s.

The new score is called SNAP-II, which includes 6 of the original 37 variables from SNAP as a measure of newborn illness severity. Birth weight, low Apgar score, and small for gestational age were added to create a 9-variable SNAP-Perinatal Extension-II (SNAPPE-II) to predict in-hospital mortality, as listed in Table 3-2. Excluded patients are babies released to the regular nursery within 24 hours, moribund infants when an explicit physician decision not to provide life support was made at the time of NICU admission, those admitted to the NICU after 48 hours since birth or after having been discharged home, and babies who died in the delivery room. The time frame for data collection was shortened from the first 24 hours to the first 12 hours after admission to the NICU in order to minimize the effect of early treatments. The discrimination of deaths from survivors by SNAPPE-II is good. SNAP-II and SNAPPE-II apply to

babies of all birth weights and are simple, rapid and accurate. However, they are more useful in estimating outcomes for a group of patients rather than a single patient [Frize M. *et al.* 2001]. New approaches should attempt to make estimates on a patient by patient basis.

Table 3- 2 SNAP-II and SNAPPE-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	< 3 rd percentile	12
Apgar score at 5 minutes	< 7	18

^a MBP= mean arterial blood pressure

3.3 About Canadian Neonatal Network (CNN) Database

3.3.1 The CNN Database and Missing Values in CNN Database

The CNN Database contains 20,488 admissions from seventeen NICUs from across Canada during the 22-month collection period (January 8, 1996–October 31, 1997). Data were collected on Day 1 (admission to the NICU), Day 3, Day 14 and Day 28 (or

discharge). Our two models used Day 1 data, which were collected within the first 12 hours in the NICU.

Similar to other databases, the CNN Database also suffers from the missing value problem. There are several reasons for missing data: [Ennett CM. 2003]

1. The data were not collected because
 - (1) it was unnecessary for the problem diagnosis;
 - (2) it was presumed normal;
 - (3) may not be done in a urgent situation;
2. The test results were missing because they were lost or contaminated.
3. Outliers. The outlier may be a data-entry error like entering “0.21” rather than “21” for the fraction of inspired oxygen; impossible values like a value below zero for heart rate; or even cases of extreme but plausible values for special cases such as a baby with a respiratory rate of 300 breaths per minute (rare but possible).

The simple and easy approach for dealing with missing data is to discard all incomplete cases, which should be with the premise that complete cases are a random sub sample adequately representing the original sample population with respect to the outcomes. For several reasons, this hypothesis may be violated. This approach also causes information loss. Sometimes missing values can be replaced with “normal” values when the data were not collected because unnecessary or the parameters were presumed normal. This approach for replacing missing values can bias results towards less sick cases [Knaus WA. *et al.* 1981]. In order to reduce any biasing effects, outliers are often simply replaced

with the appropriate minimum or maximum values [Ennett CM. 2003]. There are also many kinds of statistical imputation of missing values, such as mean imputation, multiple-draw imputation, random number generation, and nearest-neighbour algorithms. Although several methods have been commonly used, no single method is suitable for all situations. In the medical environment, the parameters used for diagnosing various problems are often different. This leads to a complex missing value problem and inconsistency in the data collection for the patient population [Pesonen E. *et al.* 1998]. Unfortunately, ANNs cannot run experiments with incomplete cases, so a new method was created for dealing with the missing data [Ennett CM. 2003].

3.3.2 Imputation of Missing Values by Integrating ANN and CBR

This approach [Ennett CM, 2003] created ANN and CBR hybrid system to impute missing data. It uses the weights of a linear ANN as the matching weights of the CBR to find the ten closest matching cases. The missing values in the NICU database were replaced with the mean of the matched cases. The hybrid approach was flexible in that no prior statistical knowledge was required of the data with missing values. The hybrid system with the imputed values had the greatest area under the Receiver Operating Characteristic (ROC) Curves and the highest sensitivity and specificity rates compared to the models developed with mean and random imputation. In addition, this model achieved the best area under the ROC curve and had good sensitivity and specificity rates compared to the hybrid models developed using the clinician-advised weights. [Ennett CM, 2003]

3.4 ANN Tool Trained with 5088 Cases from CNN Database

The ANN tool was developed by MIRG for health care professionals estimating the following outcomes for single newly admitted patients to the NICU: mortality (survival or death), the need and duration of artificial ventilation, and length of stay in the NICU. This tool extends the approach of the adult ICU ANN tool developed by our research group, which successfully used ANN to estimate ventilation duration for adult ICU patients [Frize M. et al. 1995; 2000]. The ANN tool design used a feedforward back-propagation two-layer (input and output layer) ANN, without weight-elimination. The hyperbolic tangent transfer function was selected in an application to neonatal ICU patients [Tong Y. 1999, Tong Y. 2002]. Nine SNAPPE-II variables were used as input variables. The five output variables are: Mortality, Vent0 (patient requires a ventilator or not), Vent24 (patient requires a ventilator for more or less than 24 hours), LOS7 (patient's length of stay is more or less than 7 days) and LOS28 (patient's length of stay is more or less than 28 days). The data comprised 5,088 complete patient cases from the Canadian Neonatal Network's NICU database, among which 2/3 (3,392) were randomly selected to create the training set, and with the remaining 1,696 cases used in the test set. Table 3-3 [Frize M. 2003] shows the performance of our ANNs for the NICU_ANN model in measuring sensitivity (the proportion of positive cases that are correctly identified), specificity (the proportion of negative cases that are correctly identified), and classification rate (CR, which identifies the rate at which the model correctly classifies the data into their proper categories). It is also useful to compare the CCR with the constant predictor (CP, a simple statistical tool that classifies all cases as belonging to the category with the highest training-set *a priori* probability). As can be observed in Table

Table 3- 3 Classification Performance of ANNs: Mortality, Ventilation and Length of Stay in the NICU [Frize M. 2003]

	Mortality	Vent=0 hr	Vent>=24hr	LOS<=7d	LOS>28d
Sensitivity(%)	26.6	27.8	24.4	23.7	76.9
Specificity(%)	98.0	95.3	97.1	94.3	77.5
CCR(%)	91.3	82.7	89.4	76.2	77.2
CP(%)	90.0 ^a	81.0	88.0	73.0	66.0

3-3, the CCR is systematically higher than the CP, and substantially higher for predicting long-term stay in the hospital.

3.5 Training ANN Tool with 19398 Cases from CNN Database

Recall the CNN database contained 20488 cases. Only 5088 cases had been complete. In a previous study using this database, Ennett et al. [Ennett CM. *et al.* 2001] showed that patients who had fewer missing values had a higher mortality rate than patients with more missing values. Thus the 5088 complete cases represented the population of sicker infants. Therefore, the ANN tool trained with the 5088 cases was not appropriate to be used in a population of regular patients. Fortunately, we had most missing values in the CNN database replaced so that we could use them to re-train the ANN tool in this study (refer to 3.3.2).

3.5.1 Re-training the ANN Tool

3.5.1.1 ANN Architecture

The original ANN tool was designed using the feed forward architecture trained with the back-propagation algorithm, which proved to be the best-performing network compared with various types of network architectures [Ennett CM. *et al.* 2004]. Therefore, this architecture was also employed for re-training the tool.

3.5.1.2 Preparation of Data

We trained the ANN with the entire data in the CNN database in order to eliminate the bias. Moribund babies were excluded from this study since they only received comfort measures rather than aggressive therapy [Ennett CM. 2003]. After excluding moribund infants and those missing data on the three outcome variables, as well as birth weight, small for gestational age status and Apgar score at five minutes, the number of cases remaining was 19398. The missing values for the 19398 cases were replaced as mentioned before. Table 3-4 showed the proportion of cases for the three outcomes:

Table 3- 4 Proportion of cases in the CNN database

Mort		Ventday		LOS	
Survival (-1)	18673(96.3%)	ventday<24h (-1)	12099(62.4%)	LOS<=28d (-1)	16113(83.1%)
Death (+1)	725 (3.7%)	ventday>=24h (+1)	7299(37.6%)	LOS>28d (+1)	3285(16.9)

From Table 3-4 it can be seen that the outcome “death” was 3.74%, which was low.

When a dichotomous output is highly skewed towards one outcome, the ANN typically classifies all cases as belonging to the outcome with the greatest *a priori* probability. One approach that had been shown to improve the classification performance of ANN was to selectively sample the cases with the rare outcome and create training sets with an artificially increased representation of the rare outcome. When ANN was trained with the higher percentage of rare cases, the test set performed better [Ennett CM. 2003].

Therefore, 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. The percentage of death artificially reached 10.4%.

Once the appropriate raw input data had been selected, it had to be normalized, as ANNs required data to be prepared in a certain format in order to produce accurate results. By normalization, data were scaled into an acceptable range for networks and were distributed evenly. The network with the symmetric data would learn much faster. All input and output fields must be numerical. As the hyperbolic tangent function was employed as the transfer function for our ANN training, the input values were required to be scaled between -1 and 1. This was realized by taking each input variable in a column, subtracting the mean of that column, and then dividing by three standard deviations. Also, the data file should be randomized, which was accomplished by using the random number generator to generate a value for each row and then sorting the column with the random numbers. After the process of normalization, the data were randomly divided into training set and test set. The training set contained 13899 cases, which were the two-thirds of the 20848 cases, while the remaining one-third went to the test set, which were 6949 cases. The training set was used for the adjustment of the weights during training, while the test set functioned as the verification set, which was used to report the performance of the ANN. A validation set should be added to check the performance of the ANN after training and testing. The EPIC set acts as a validation set for this project. Then the data were ready to start the training process.

3.5.1.3 ANN Implementation Process

The ANN training process was conducted using nine SNAPPE-II variables as inputs and three variables as outputs from the CNN database. Different ANN structures were employed for this experiment, including two-layer and three-layer networks, with and

without weight-elimination. The performance was evaluated based on the results of ASE (average squared error), CCR, sensitivity, specificity and the ROC curves.

Before training commenced, the connection weights for the input layer and hidden layer were set to random values which were between -1 and 1, and -0.1 and 0.1 respectively [Ennett CM. *et al.* 2004]. The initial values of the following network parameters needed to be set each time when training started: learning rate, learning rate increment, learning rate decrement, momentum, error ratio, weight-decay constant and weight-elimination scale factor. The automated parameter adjustment method was employed to calibrate each ANN parameter for the achievement of the ANN's optimal performance for the training set, which was defined as the maximum of CCR and the highest sensitivity and specificity [Frize M. 2000, Ennett CM. 2004]. Then the results were tested in the test set to see if the training process was appropriated.

First, the network was set to the two-layer structure without weight elimination. The stopping criterion was set to the CCR, the ASE and the log-sensitivity index respectively in order to get the best results. After the optimal results were achieved, the weight-elimination cost function was switched "ON", which allowing network pruning by eliminating variables that offered little or no assistance in predicting the correct outcome. Also, by limiting the size of the connection weights, thereby penalizing large weights, this cost function resulted in a more stable network, because the output had less variance. Then the network was set to the three-layer structure without weight elimination. The three stopping criteria were also employed. The network was trained by starting with one

hidden node, and one more node was added each time until the network's performance began to decline. Then the weight-elimination cost function was employed to see whether the result had improved or not.

3.5.2 ANN Training Results

Recall we mentioned before, ANN was trained and tested for the best performance based on the results of ASE (average squared error), CCR, sensitivity, specificity and the ROC curves. There was always a trade off when choosing the best value from the training results. Higher rates of sensitivity and specificity were more desirable for our ANN tool, while the CCR value were preferred to be higher than that of the CP. Comparing sensitivity and specificity, higher sensitivity rate was chosen with the premise that the specificity rate would not decrease by much. Table 3-5 displays the classification performances of our ANNs with respect to measuring sensitivity, specificity, CCR and ROC for the outcome variable mortality (Mort) and Length of Stay in hospital (LOS).

Table 3- 5 Classification performance for Mort and LOS

	Mort	LOS
Sensitivity(%)	70.2	71
Specificity(%)	93	85.9
CCR(%)	90.6	83.4
CP(%)	89.5	83.1
ROC	0.88	0.84

The best results for the variable "Mort" and "LOS" were obtained by training two-layered network, without weight elimination, and with log-sensitivity index as the stopping criterion. ROC curves are shown below in figure 3-1 and 3-2.

Figure 3- 1 ROC curve for “Mort”

Figure 3- 2 ROC curve for “Los”

The best results for the output variable “Ventday” (Table 3-6) were obtained by training three-layered network (five hidden nodes), with weight elimination, and with ASE as the stopping criterion. ROC curves are shown below in figure 3-3. Table 3-6 displays the best results for Ventday with two-layered and three-layered network training. Figures 3-3 and 3-4 show that ROC curve of Ventday with three-layered training is better than that of with two- layered training.

Table 3- 6 Classification performance for Ventday

	Ventday (3-layer)	Ventday (2-layer)
Sensitivity(%)	76.7	63.4
Specificity(%)	90.6	89
CR(%)	84.8	78.8
CP(%)	58.8	58.5
ROC	0.90	0.82

Figure 3- 3 ROC curve for “Ventday” (3-layer)

Figure 3- 4 ROC curve for “Ventday” (2-layer)

3.5.3 ANN Tool

Weights corresponding to the nine input variables and the biases (Table 3-7) extracted from the best training results of each output variable were inputted into the prototype ANN tool. Recall that the best results for ‘Mort’ and “LOS” were obtained by training the two- layered network. Therefore, there was one connection weight for each input variables.

Table 3- 7 Weights and biases used in the ANN tool

Mort	W1: -0.5165 -0.4617 -1.1620 -0.6265 -0.5371 0.4556 -0.7572 -1.2489 0.5545
	B1: -1.5426
LOS	W1: -1.0763 0.4753 0.2443 -0.0421 0.3539 -5.1643 -4.2307 0.3216 -0.2704
	B1: -1.8562
Ventday	W1: 1.1346 0.6258 -0.2125 0.8953 1.3363 -0.5740 2.0272 0.1894 0.4203
	-0.2803 0.2077 6.3298 -0.4324 -0.4999 -0.2496 -0.0286 -0.1241 0.2437
	-0.0322 0.0646 -1.2802 -0.3091 -0.1587 -0.6026 -0.3565 -1.1130 -1.2626
	-0.1402 0.0214 6.5308 0.8089 0.1046 0.3177 -0.4898 -0.3382 -0.4985
	0.2519 -0.3946 0.9423 -1.1088 -2.4563 -1.1095 1.2404 0.9782 -0.1999
	B1: 1.4563 -0.7729 0.6479 0.5393 1.0060
	W2: -1.1363 1.3009 0.5950 -1.6591 -1.4425
	B2: 2.7385

W: weight

B: bias

The best results for the “Ventday” were reached through three-layered network training with five hidden nodes. W1 represents the connection weights from the input layer (nine input variables) to the hidden layer (5 hidden nodes). Each of the nine input variables was connected with each of the five hidden nodes. Therefore, there are five connection weights corresponding to each input variable. W2 refers to the connection weights from the hidden layer to the output layer, whereas each of the five hidden nodes was connected with the one output variable. So there was one connection weight corresponding to each hidden node (refer to figures 2-2 and 2-3).

Figure 3-5 displays the interface of the ANN tool. Nine input variables are on the left side,

Health Care Professional Screen

MIRG ANN Decision Aid Prototype for Clinical Trial

Click 'New Patient' to initialize

Patient Characteristics		Patient Outcomes	
New Patient	Patient ID 59 ?	Mortality -0.99991 ?	
LBLLOODP 57.299 ?		Vent0 0 ?	
LTEMPF (deg F) 97.88 ?	SEIZURE 1 ?	Vent24 -0.98209 ?	
PO2FI02R 4.8 ?	BTHWT (grams) 1011 ?	LOS 7 0 ?	
LURINE (CCs) 8.4 ?	SGA (0-1) 0 ?	LOS 28 -0.18924 ?	
LSERUM (pH level) 7.34 ?	APGAR5 (0-10) 5 ?	All	
		Save Record	

Figure 3- 5 Interface of the ANN tool

while five output variables on the right side. For each new patient, click the 'New Patient' button, and then type in the patient ID and nine input variables. For the outcomes, just click the corresponding button for each of the five outcomes and the outcome you want will be displayed beside the button. If you click the button 'ALL', the five outcomes will be displayed at the same time. If you click the 'Save Record' button, all of the input and output information will be saved for future use. The ANN tool calculates outcomes based on the values of weights and biases. The value of the outcomes is either -1 or +1, because all values above zero are defined +1 and all values below zero are defined as -1. Intermediate values have no specific meaning at this time.

Chapter 4 Methodology

A pilot study was designed to do a preliminary evaluation on our ANN tool and to prepare work for a future clinical trial. The design of the study was based on the evaluation methods for medical expert systems discussed in chapter 2. As the ANN tool was at the prototype stage, the second stage of evaluation of medical expert systems [Wyatt J. *et al.* 1990] was applied. The study included an attitude test and a laboratory experiment. The attitude test aimed at measuring attitudes of users and collecting their suggestion for our ANN tool. Laboratory experiments were designed to test the performance of the prediction to assess whether the predictions of our ANN tools were higher than, or equal to, or lower than those of the physicians. The information gathered and analyzed was both qualitative and quantitative. As for the limited amount of data sources and small amount of time available for physicians to spend on the experiment, the data size for quantitative analysis was limited.

Before the study began, a proposal [appendix 4] for the project, with the title “Comparison of Predictions by an Artificial Neural Networks Tool and Physicians of Outcomes of Neonatal Intensive Care Unit Cases”, was created and submitted to the CHEO Research Ethics Review Committee. The proposal included the background and research objectives and the methodology to be used in the study. Since the research protocol involved minimal risk, the proposal could be considered under the rubric of an expedited review. Therefore, this project was approved quite rapidly.

Informed consent had to be obtained from each physician before his/her participation in the study. A consent form was created which included the following information: the purpose, procedures, benefits, discomforts, risks and precautions associated with this study. It also described physicians' right to refuse to participate or withdraw from the study at any time. In order to decide whether physicians wished to participate in this research study, they were required to understand enough about its risks and benefits to be able to make an informed decision. This is known as the informed consent process. The following sections provide details on the design and implantation of the two studies.

4.1 Pilot Study 1–Attitude Test

4.1.1 Design of the Attitude Test

The attitude test was designed in the form of a questionnaire consisting of three parts. Part 1 was a set of close-ended questions with 5-point Likert scale, which aimed to measure the users' attitude towards using the ANN tool. A 5-point Likert scale ranging from strongly agree to strongly disagree, including a neutral mid-point was used for each question. The strongly agree was labeled as “very likely” while the strongly disagree as “very unlikely”. The mid-point was labeled as “uncertain”.

The first question assumed a tool estimation that is similar to the physician's estimation. The second question assumed a tool estimation that is significantly different from the physician's estimation. Based on these two situations, users were asked whether they would like to use the ANN tool to assist them in their diagnosis and treatment of patients. In the next question, “If your estimation of probability of death for a patient were

significantly different from the ANN tool's estimation, how likely would you be to accept each prediction?", the output variable Mort was chosen to see the likelihood of users' willingness to use this tool. The reason for choosing the Mort rather than other two output variables LOS and Ventday was that the estimate of patient's survival or death was more important to physicians. The next question asked users whether they thought the ANN tool would be helpful when estimating the following three outcomes: Mort, Ventday, and LOS. Part 1 of the questionnaire [see appendix 2] also contained questions on whether the information provided by ANN tools would help users in estimating the risk of complications in patients and whether the tools would speed up the decision-making process.

Part 2 of the questionnaire was structured as open-ended questions. This part aimed at gathering information for the improvement of the ANN tool. It allowed users to express their own opinions and suggestions freely. The first question required the physicians to rank the importance of the variables used in their process of estimation. The next question asked users what other patient data would be important in addition to the input variables used for ANN tools in predicting outcomes. The two questions were designed to find other important input variables for our system, and also to assess the weight of each variable physicians use when they estimate outcomes for their patients. We can compare the difference of the weight given each variable by physicians, with those set by the ANN tool, and see if we can improve our tool's performance. The questionnaire also allowed users to comment on the ANN tool, therefore providing suggestions for its improvement of our tool.

Part 3 of the questionnaire collected demographic information on the participants—the neonatologists in the NICU of CHEO. In the medical field, experience is important for physicians to diagnose and treat patient. Therefore, questions on physicians' medical training location, years of practice in neonatology and in other medical fields were asked. Due to the small sample size (small number of neonatologists at CHEO) we can not make a quantitative statistical analysis. However, this information was valuable for a qualitative analysis.

4.1.2 Implementation Process of the Attitude Test

4.1.2.1 Subjects

The subjects were six neonatologists in the NICU at the CHEO and at the Ottawa Hospital-General Campus, who wished to participate in this study. Four were male and two were female. Three doctors received their medical education in Canada, while two obtained theirs in both Canada and United Kingdom, and one in both India and Canada. For the years of practice as a neonatologist, three had more than 10 year experience, while the other three had less than 5 years.

4.1.2.2 Procedure

The consent form was given to each neonatologist who worked in NICU at CHEO and Ottawa Hospital-General Campus to invite them to take part in the study. Six neonatologists agreed to participate. The questionnaire was sent to each of them, together with the cases used in pilot study 2.

4.2 Pilot Study 2: Laboratory Experiment

4.2.1 Design of the Experiment

This experiment was designed to test the performance of the ANN tool compared with those of physicians. The data used for this project were from “CNN EPIC 2002 CHEO v3_CHII” database (EPIC database), which contained fifty nine patient cases, with the nine variables needed for the ANN tool, information for physicians to make predictions, and real outcomes. All fifty nine cases were tested with the ANN tool and also by physicians. Then the performance of the ANN tool and of physicians was compared.

4.2.2 Implementation Process of the Experiment

4.2.2.1 Data Requirements:

Data used in this experiment must satisfy the following requirements:

1. Containing nine input variables of ANN tool;
2. Containing three outputs of ANN tool;
3. Containing information needed for physicians to do predictions.

Three outputs (mortality, requires a ventilator more or less than 24 hours, length of stay is more or less than 28 days) were chosen from five possible outputs of the ANN tool to do the comparison with physicians’ predictions. This was suggested by Dr. Walker, an expert in neonatology at CHEO, because of the limited amount of time available of the physicians participating in this study.

Dr. Walker also suggested a division of information into five parts for physicians to estimate outcomes, as shown in Table 4-1.

Table 4- 1 Patient's information needed for physicians

Admission Information	SNAP
Case number Gender Birth weight Gestational age (Pediatric estimate) Gestational age (Obstetric estimate) Apgar Score @ 1 minute Apgar Score @ 5 minute Apgar Score @ 10 minute Births this pregnancy Admission weight	Blood pressure low (bplow) Respiratory rate (resprate) Temperature Serum pH Lowest pO ₂ -FiO ₂ (lpO ₂ fiO ₂) Lowest pO ₂ -MAWP (lpO ₂ mawp) Seizures Urine output CC's (ml/Kg.H)
Mother's information	NTISS
Mother's age Gravida Total abortions Labour initiation type Delivery type Maternal hypertension Presentation Prenatal care Antenatal corticosteroid Diabetes Chorioamnionitis Tocolysis	Supplemental O ₂ CPAP Mechanical ventilation Ventilation with relaxant (ventrlx) High frequency ventilation (hifqvent) Surfactant Nitric oxide (NO)
Diagnosis/procedures	
PDA RDS Pneumothorax	

4.2.2.2 Data Sources:

Available databases were the CNN database and the EPIC database. As the former does not contain enough information needed for neonatologists to perform estimations, it could not be used in this project.

EPIC database contained 59 real patients' data collected at CHEO in the year 2002 for the Canadian Neonatal Network EPIC (Evidence-based Practice Identification and Change) study. All of the 59 cases were premature infants (gestational age less than or equal to 32 weeks) and were transferred from other hospitals. This database satisfied all the requirements for this study.

4.2.2.3 Extracting Data from Database and Data Preparation:

The EPIC database was created in Microsoft Access, which contained 53 tables. Each table was a collection of data about a specific topic. The information needed was extracted from different tables by creating queries. Then data transformation and normalization were performed.

(1) Data for ANN Tool:

Among nine input variables (Lowest blood pressure, Lowest temperature, Lowest PO₂/FiO₂ ratio, Lowest serum pH, Multiple seizures, Lowest urine output, Birth weight, Small for gestational age, APGAR score at 5 minutes), five of them (Lowest blood pressure, Lowest serum pH, Multiple seizure, Birth weight, APGAR score at 5 minutes) were directly extracted from the tables.

'Lowest temperature' was transferred from Celsius degrees to Fahrenheit degree according to the requirement of the ANN tool.

The variable Lowest PO_2/FiO_2 ratio was not in the database. The variable lpO_2fiO in this database was only in fact the FiO_2 part of the lowest PO_2/FiO_2 ratio (i.e., if the value for variable lpO_2fiO is 80 in the database, it means at the point when the lowest PO_2 was obtained, the FiO_2 was 80%). The other part of the ratio was the lowest PO_2 which was lpO_2lpO_2 in the database. Then the Lowest PO_2/FiO_2 ratio could be obtained by dividing lpO_2lpO_2 by lpO_2fiO .

In the database, the lowest urine output was the variable $urinecc$, recording the total urine output values during the 12-hour period, while in the ANN tool, this variable was in the units of mL/Kg.H. Therefore, $urinecc$ values in the database were divided by the baby's weight (in kg), and then again divided by 12 in order to match the units of the tool.

SGA (small for gestational age) means the baby weighed less than the 3rd percentile for its gestational age. It is determined by placing the baby's birth weight and gestational age on a chart. It was not a 'data point' so it was not in the tables. A tool on the web [SGA tool online] was used to determine whether a baby is SGA. The tool gave out the 3rd percentile for each gestational age; therefore, if the baby's weight was less than this number for the gestational age, he/she was labeled as SGA. The gestational age for each baby in the EPIC database was saved as $obsga$ and $pedga$, which was on obstetric estimate of gestational age and a pediatric estimate of gestational age respectively. The obstetric estimate is usually considered the more accurate as it is based on early

ultrasound data and other assessments. But if there was not an early ultrasound or other reliable estimate during pregnancy, then the estimated gestational age from the baby's physical examination at birth was used, which was less accurate. So a value in obsga was taken even if there was a value in pedga. If there was no value in obsga (missing value) then the value in pedga was adopted.

For nine input variables, five of them had complete values for 59 cases. There was one missing value in APGAR, two in Lowest urine output, fifteen in Lowest serum pH and twenty in Lowest PO_2/FiO_2 ratio. The missing values must be replaced because the tool required nine complete input values for each case. Dr. Walker suggested that these missing values could be replaced with normal values since it was clear that the missing values were not available because the tests were deemed unnecessary (i.e., results would likely have been normal.). The normal value for each variable could be obtained from the SNAP scoring form, where the normal value is represented under the scaling score of '0'. For the three output variables, only LOS was not available in the database, but it could be obtained by calculating the time between the admission date and the discharge date of the patient. Since some patients were transferred to a second or even a third hospital and there were no records of discharge date, only 30 of the 59 cases had the outcome of LOS.

(2) Data for Physicians:

In total, 40 fields of patient information were provided to neonatologists, which included patient's admission information, mother's information, SNAP, NTISS (the Neonatal Therapeutic Intervention Scoring System) and admission diagnosis. All of the information was collected within 12 hours of admission.

The missing values were not replaced with normal values since the data were exactly in a form encountered by physicians in real practice.

Many parameters were coded in the database instead of real values; an example is: for the variable 'delivery type', 1 represented that the delivery type was vaginal, 2 was caesarean, and 3 was unknown type. When preparing data for physicians' prediction, these numbers were converted to their real meanings.

For ethical and privacy considerations, some information could not be extracted directly from the database. For example, a mother's date of birth could not be given out as this might identify her. Instead, her age in years at the date of the baby's birth was obtained by calculating the mother's date of birth and her baby's date of birth.

Also, a scenario was created for each of the fifty nine cases based on the information in the database. It included the infant's general information such as gender, gestational age, weight at birth, basic physiological exams, current health status, and the mother's information such as age, health condition during the pregnancy, medication used, and delivery type. By presenting a scenario with other detailed information (forty fields in Table 4-1) of a patient, physicians would understand the patient's context better than just giving them a list of numbers or words to do the predictions.

When all the data were gathered, the fifty nine cases were divided evenly into six groups. Each group contained the cases which had roughly the same proportion of the three

dichotomous outcomes. As a result, physicians could do the predictions on roughly similar cases.

4.2.2.4 Predictions of Physicians

Each physician was given a questionnaire and ten cases. The results are discussed in the next chapter.

4.2.2.5 Predictions of ANN Tool

The nine input variables for each case were put into the ANN tool and the predictions for each of the fifty nine cases were obtained. The predicted results were compared with the real results for each patient, based on which the sensitivity, specificity and CCR were calculated for the assessment of the tool. The results are discussed in the next chapter.

Chapter 5 Results and Discussion

5.1 Results and Analysis for the Attitude Test

Part 1 of the questionnaire applied a 5-point Likert scale instrument, which is commonly used for quantitative analysis. This pilot study performed simple calculations for quantitative analysis due to the small sample size, while the qualitative analysis was also conducted at the same time. For each question, different scores were assigned to five

Table 5- 1 Responses to the part 1of the questionnaire

Questions	number of answers for each category					Ave
	2 (Very likely	1	0	-1	-2) Very unlikely)	
1. Would you use the ANN tool to assist your diagnosis and treatment of patients if the tool's outcome predictions were similar to your prediction as a physician?	2	1	2	1		0.67
2. Would you use the ANN tool to assist your diagnosis and treatment of patients if the tool's outcome predictions were significantly different from your prediction?		2	3		1	0
3. If your estimation of probability of death for a patient were significantly different from the ANN tool's estimation, how likely would you be to accept each prediction (below)?						
Likely to accept your estimate		6				1
Likely to accept ANN's estimate			4	2		-0.33
4. Do you think the information (ANN tool's outcome prediction) would help you in estimating mortality in terms of short-term survival (until discharge)?	1	3	2			0.83
5. Do you think the information would help you in estimating the likely length of time a ventilator would be needed?		4	1		1	0.33
6. Do you think the information would help you in estimating the patient's length of stay in NICU?	1	4		1		0.83
7. Do you think the information would help you in estimating the risk of complications (e.g., IVH, PVL) in the patient?		3	2		1	0.17
8. Do you think the ANN tool would speed up your decision-making process?		2	1	3		-0.17

answer choices. Score 2 was assigned to the most positive answer, 0 to the neutral answer, while -2 to the most negative one. The average category was calculated and used as criteria. For example: in question 1, the average score is $(2*2 + 1*1 + 2*0 + 1*(-1)) / 6 = 0.67$. A score of 0 was deemed to be the cut-off point, above which the opinions were deemed to be positive; below this point, their opinions were deemed to be negative.

The difference between the first two questions is that: the first one supposed that the outcome predictions of the tool were similar to physicians' predictions; the second one supposed that predictions by the tool and physicians were significantly different. The first question received an average score of 0.67, while the second one received an average score of 0. Thus, the ANN tool seemed to be welcomed as confirmation of a physician's diagnosis, and recommended treatment of patients, when the tool's predictions were similar to those of physicians, which means the tool can strengthen their decisions. If the ANN's predictions were significantly different, then some physicians' weakly positive attitudes outweighed the strongly negative attitude by two to one, while three physicians chose the neutral answer. From the answers, we can see that some physicians are not quite sure whether to use the tool or not when the tool's predicted outcomes are significantly different from theirs.

Question 3 in Table 5-1, for a certain outcome, "Mortality", when the estimation result of ANN tool was significantly different from that of physicians, there were two possible answers. This question was designed to let physicians choose one of the two choices, which was either "accept physician's own estimation", or "accept ANN's estimation". It

turned out that all the physicians selected both answers. For the first answer 'Likely to accept your estimation', all of them chose the 'likely' choice instead of the choice 'very likely'. This reflected the physicians' attitude that even though they would like to accept their own estimation, it is not so strong to make them refuse other opinions or suggestions. This attitude was also shown when they chose the second answer 'Likely to accept ANN's estimation'. Four of them chose the neutral answer, and the other two chose 'unlikely'. None of them chose 'very unlikely'. From these answers, we can see that most of the physicians do not reject the ANN tool. They chose the neutral answer perhaps because they are not yet familiar with the tool. In the future, if the tool is shown to have an acceptable performance, physicians would likely use the tool for assistance, even though the tool's predictions may be different from their own.

Next, when asking whether the information provided by the ANN tool would help in estimating particular outcome such as mortality, ventday, length of stay and the risk of complications, the positive answers outnumbered the negative answers by 16 to 3, which demonstrated that most physicians believed the ANN tool would help them in estimating these outcomes. For question 8 "do you think the ANN tool would speed up your decision-making process?", two physicians answered 'likely', one said 'uncertain', while the other three answered 'unlikely'. From their answers we found that most physicians did not believe that the tool would speed up the decision-making process.

To sum up, of the answers to the first eight questions for each physician, five physicians showed a positive attitude towards using the tool, while one demonstrated a negative

attitude. Most negative answers came from this physician. Also, by analyzing individual answers, it seems that physicians with relatively more experience were more willing to use the tool. Furthermore, physicians with more experience were more likely to choose either positive or negative answers, but not the neutral one, as compared with physicians with less experience. Based on the overall average category values (see Table 5-1), we concluded that physicians had mainly positive responses for all questions concerning whether they would like to use the ANN tool as an assistant, with the exception of question 2 – where ANN predictions differed significantly from the physician's prediction – the attitudes were mainly neutral.

Although six physicians' attitude on using the ANN tool represents most of the physicians' attitude in the NICU at CHEO (total of nine physicians), the sample size is small. However, being a medical doctor myself in my former training, I believe that if the ANN tool's performance is similar to or better than the physicians' performance on prediction, physicians will likely accept the tool for assistance. It is possible, therefore, that many other physicians' attitude in NICU across the country would reflect the attitudes found at CHEO.

The second part of the questionnaire started from question 9[refer to appendix 2]. This part focused on gathering opinions and suggestions from physicians to improve the system. Question 9 listed fifteen variables used in either the ANN or in a new neonatal mortality model developed by Ennett [2003]. Physicians were asked to rank, from 1 to 5, the five most important variables they would use to estimate outcomes. The results of

rank by six physicians are shown in Table 5-2. In order to rank all of the variables used by six physicians, we also assigned a score for each variable according to its rank position given by physicians. For example, variables in “Rank 1” column (see Table 5-2) were assigned 5 scores, in “Rank 2” were 4 scores, and so on. The rank of the total variables could then be obtained by summing the scores.

In the table, each physician has a different variable combination. However they all selected birth weight (Bthwt) as one of the variables to estimate outcomes. Furthermore,

Table 5- 2 Rank of variables by each physician

	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5
Physician1	Bthwt	Po2fio2r	Lserum	Sga	gestational age
Physician2	gestational age	Bthwt	apgar5	Po2fio2r	Lserum
Physician3	Bthwt	gestational age	Sga	Ltempf	Lbloodp
Physician4	Bthwt	Specific diagnoses	Ventilation type/needs	Po2fio2r	apgar5
Physician5	gestational age	Bthwt	IVH grade 3/4	Congenital anomalies	Lbloodp
Physician6	Bthwt, Sga, Lserum, Seizure	apgar5, Lbloodp, Ltempf, Hrespr, Po2fio2r, Hpco2, Lgluc, Lplt	Hbloodp, Hsodium, Lurine		

Bthwt (birth weight)

Sga (whether small for gestational age)

apgar5 (Apgar score at 5 minutes)

Hblood (highest blood pressure in 1st 12 hours)

Lbloodp (lowest blood pressure in 1st 12 hours)

Ltempf (lowest temperature in 1st 12 hours)

Hrespr (highest respiratory rate in 1st 12 hours)

Lserum (lowest serum pH in 1st 12 hours)

Po2fio2r (lowest pO2:FiO2 ratio in 1st 12 hours)

Hpco2 (highest pCO2 in 1st 12 hours)

Hsodium (highest serum Na in 1st 12 hours)

Lgluc (lowest blood glucose in 1st 12 hours)

Lplt (lowest platelet count in 1st 12 hours)

Lurine (urine output in 1st 12 hours)

Seizure (whether multiple seizures)

four of them ranked Bthwt as the most important variable, another two ranked it as second in importance. Therefore, Bthwt received the highest score. We could reach the conclusion that Bthwt was the most important variable when predicting outcomes by physicians. Table 5-3 shows the rank of the variables chosen by six physicians for their top five rankings in outcome estimations.

Table 5- 3 Variables chosen by physicians and the rank

Link#	Variables	Rank	Scores	used by # of physicians
1	Bthwt	1	28	6
2	Po2fio2r	2	16	4
3	gestational age	3	15	4
4	Sga	4	10	3
5	Lserum	5	9	3
6	apgar5	6	8	3
7	Lbloodp	7	6	3
8	Ltempf	8	6	2
9	seizure	9	5	1
10	Hrespr	10	4	1
11	Lgluc	10	4	1
12	Lplt	10	4	1
13	Specific diagnoses	10	4	1
14	Hpco2	10	4	1
15	Lurine	11	3	1
16	Hsodium	11	3	1
17	Hbloodp	11	3	1
18	IVH grade 3/4	11	3	1
19	Ventilation type/needs	11	3	1
20	Congenital anomalies	12	2	1

Note: variables in bold are those used in the ANN tool

In Table 5-3, twenty variables were used by physicians. In addition to the fifteen variables provided, five variables (Link #3, 13, 18, 19, 20) were added. These were variables that some of the physicians considered to be important in estimating outcomes. Among them gestational age had high scores, which demonstrated that this variable was very important for physicians in predicting outcomes. Note that eight out of nine variables used by ANN tool were ranked high in the table, which demonstrates that variables used by physicians and by the ANN tool are similar.

Question 10 concerned variables other than the fifteen listed in question 9. These were variables that physicians may have added in addition to their ranking of the top five, which could include other patient data that they thought might be more important for making diagnosis, deciding treatment and management of the patient, or predicting prognosis. Results are given in Table 5-4. One physician mentioned the variable “gestational age”, which was not on his/her top five ranking list. Recall that there were four physicians using this variable as one of the top five variables when predicting

Table 5- 4 Answers for question 10

Physician	Answer
1	Admitting diagnosis, congenital anomalies
2	Oxygen index, Sarnat score
3	Head ultrasound result
4	Gestational age
5	Unlimited + extensive need + + info beyond 1 st 12 hours. Needs very hugely depending on GA and overall clinical picture
6	Maternal Medical History and Infectious status and Morbidity prior to delivery

patient's outcomes. Gestational age was also considered by five out of six physicians when predicting outcomes; however, this variable was not chosen for training the ANN tool. In our future work, this variable can be added to ANN to see if the performance is improved. The variables "congenital anomalies" and "diagnosis" were given by two physicians (Table 5-2 and Table 5-4), which suggested that some physicians might consider them as factors. Physician 5 felt that much more information was needed in order to make a diagnosis and prediction. This is true in clinical practice. However, our tool is focused on the first 12 hours' information and would help physicians to make decisions soon after patients' admission. Of course with the lapse of time and treatment, patients may get better or get worse. Nevertheless, this is not within the scope of what the ANN tool attempts to consider as inputs. Also, there are some medical data used by physicians that can not be measured quantitatively, such as admitting diagnosis, congenital anomalies, head ultrasound result, maternal medical history and morbidity prior to delivery (in Table 5-4). These variables need to be transformed into numerical data before they can be used in the ANNs.

Responses to question 11 "Other comments about the ANN tool (E.g., Would **outcomes** other than those in Q.4-7 be useful for clinical decision-making? In your opinion, what are the most positive **benefits** of the system? Do you see the possibility of anyone **misusing** the system? Any other comments?)." were displayed in Table 5-5. Five out of six physicians offered their comments.

Table 5- 5 Answers for question 11

physician	Answer
1	Long term outcome predictions (e.g., neurodevelopmental) would be very useful for counseling parents and making decisions. However, the outcomes in q 4-7 alone would certainly assist counseling. As long as health care team members give all information (including these predictions) to parents with full support and interpretation, there's little chance of misuse, but predictions should never be given alone as raw data without professional support.....
2	I see the ANN tool as an assistance to "solidify" my predictions for certain outcomes, especially in cases which we see less often. It may replace colleagues' opinions in the middle of the night but does not replace my own judgment. The cases provided here are not rare and a bit hard to understand at times. Would not see great value in ANN for regular premature infants having regular courses of disease.....
3	Head ultrasound results are very important. Benefits: quicker access to statistics, risk assessment, Draw backs: Risk figures could be used as actual figures, ignoring the individual variations. Yes, I do think that misuse is possible, and that decisions may or may not be made based on past experience / results. How can the system incorporate changes in medical management which may alter outcomes? An individual physician's experience can sometimes be important in survival vs non-survival, despite the odds / risks.
4	At this point don't know enough about it to see benefits. Have significant concerns about its ability to predict.
5	Able to share the outcome profile with the team and parents.

5.2 Results and Analysis for the Lab Experiment

5.2.1 Results and Analysis for the ANN Tool

Table 5-6 displays the real case numbers for each outcome in EPIC database and the correct prediction case numbers for each outcome by the ANN tool. Table 5-7 shows the classification performance of the ANN tool (training results) and its corresponding prediction rates for 59 cases in the EPIC database. From Table 5-6, we see that for the Mortality, the sensitivity is 25% for EPIC database, which is distinctly different from the training result of the tool, 70.2%. Also, the same thing happened in the specificity of the Ventday (trained with 3-layered network). For the ventday (trained with 2-layered network) and the LOS, the differences are not so obvious. For both Ventday (3-layer) and Ventday 2-layer), the CCRs are smaller than the CPs when predicting outcomes using

Table 5- 6 Correct prediction case # of ANN tool for EPIC

	Mort		Ventday		LOS	
	Death (+1)	Survival (-1)	Ventday>=24h (+1)	Ventday<24h (-1)	LOS>28 d (+1)	LOS<=28 d (-1)
Real outcome	8	51	45	14	19	11
ANN Tool's outcome	2	50	39 (3-layer) 18 (2-layer)	0 (3-layer) 14 (2-layer)	12	7

Table 5- 7 Classification performance of ANN tool (best testing results from training ANN using CNN database) & its prediction rates for EPIC database

	Mort		Venday(3-layer)		Venday(2-layer)		LOS	
	Tool (CNN)	Predict (EPIC)	Tool (CNN)	predict (EPIC)	Tool (CNN)	predict (EPIC)	Tool (CNN)	Predict (EPIC)
Sen (%)	70.2	25	76.7	87	63.4	40	71	63.2
Spec (%)	93	98	90.6	0	89	100	85.9	63.6
CCR (%)	90.6	88.1	84.8	66.1	78.8	54.2	83.4	63.3
CP (%)	89.5	86.4	58.8	76.3	58.5	76.3	83.1	63.3

the EPIC database, which are contrary to the results of ANN training. By analyzing the databases, we found some differences between the CNN database, which was used training the ANN tool, and the EPIC database, which was used to evaluate the tool. First, recall that in the CNN database, for each outcome variable, negative cases were far more numerous than positive cases (see Table 3-4). However, in the EPIC database (see Table 5-8), for the variable Ventday and LOS, negative cases were less than positive cases.

Table 5- 8 Case proportions of EPIC database

Mort		Ventday		LOS	
Survival (-1)	51(86.4%)	ventday<24h (-1)	14(23.7%)	LOS<=28d (-1)	11(36.6%)
Death (+1)	8 (13.6%)	ventday>=24h (+1)	45(76.3%)	LOS>28d (+1)	19(63.3)

The death rate in the CNN is 3.7%, while in the EPIC is 13.6%. These facts would suggest that the patients in the EPIC database were sicker than those in the CNN database, since more patients needed artificial ventilations and stayed longer in hospital. As well, death rates are higher in the EPIC database. In addition, the patients in the EPIC database were all premature infants, while the CNN database combined premature and normal infants. This also indicates that the population in the two databases have differences.

After examining the statistical data – means and standard deviations, we found some interesting points. From Table 5-9, we discovered that the mathematical mean and

standard deviation for the variable “po2fio2r”, “lurine” and “bthwt” in the EPIC database differed from the CNN database.

Table 5- 9 statistical values from EPIC, test and training datasets

		lbloodp	ltempf	po2fio2r	lurine	lserum	seizure	bthwt	apgar5	sga
EPIC	Mean	39.1	97.3	3.1	4.4	7.3	1.0	1192.9	7.4	0.1
	Std	13.9	1.2	1.7	3.3	0.06	0	466.2	1.6	0.3
Test Set	Mean	37.4	97.4	1.9	0.9	7.3	1.0	2383.3	7.8	0.1
	Std	9.4	1.3	0.8	0.6	0.1	0.3	1042.6	1.8	0.2
Train Set	Mean	37.5	97.4	1.9	0.9	7.3	1.0	2384.8	7.9	0
	Std	9.3	1.2	0.8	0.6	0.1	0.3	1034.2	1.8	0.2

Take closer look, comparisons between EPIC, Test and Training Sets using two tailed, unequal variance *t*-test were performed (Tab. 5-10). The probabilities of a higher value of the *t*-statistic under the assumption that the data sets are samples from populations with the same mean were calculated using Microsoft Excel. The *t*-test showed significant difference ($p < 0.05$) between the “po2fio2r”, “lurine”, “seizure”, and “bthwt” values for the EPIC and Test/Training Set. As expected, there is no significant difference between Test and Training sets ($p > 0.05$). This test is simply to point out that the EPIC data set is somewhat different than the main CNN database used in our experiments.

Table 5- 10 Probability of t-Test for EPIC, Test and Training Sets

	lbloodp	ltempf	po2fio2r	lurine	lserum	seizure	bthwt	apgar5	sga
EPIC vs Test	0.35	0.45	0.00	0.00	0.22	0.00	0.00	0.08	0.59
EPIC vs Train	0.38	0.41	0.00	0.00	0.26	0.00	0.00	0.05	0.57
Test vs Train	0.48	0.59	0.98	0.61	0.57	0.28	0.95	0.13	0.67

Next, we attempted to gauge the importance of the above four input variables when predicting outcomes. The nine input variables for Mort, Ventday and LOS of the ANN

tool were ranked according to their weights. The higher the absolute value of the weight, the more contribution the variable makes. Table 5-11 displays the rank results. Note that in Table 5-11 the ranking of the nine variables for Ventday pertains to Ventday(2-layer) and not Ventday(3-layer). The rank of Ventday(3-layer) could not be conducted, because in the 3-layered network training, the input and output relationship is not linear.

Therefore, it would be difficult to use the input weights of variables to estimate the proportion of contribution each variable makes to the outcome prediction.

In Table 5-11, the three variables (in bold) are ranked 2, 3, 4, and 9 for Mort; 1, 2, 8, and 9 for LOS; and 2, 3, 5, and 6 for Ventday. Their rank positions indicate that the three variables play important roles in predicting outcomes. Thus, the variations in the input values of these four variables may result in changes in prediction performance.

Table 5- 11 ANN tool variable weights

mort			LOS			Ventday(2-layer)		
rank	variable	weight	Rank	variable	weight	rank	variable	weight
1	apgar5	-1.2489	1	seizure	-5.1643	1	apgar5	-1.3014
2	po2fio2r	-1.1620	2	bthwt	-4.2307	2	po2fio2r	-1.2917
3	bthwt	-0.7572	3	lbloodp	-1.0763	3	bthwt	-1.0847
4	lurine	-0.6265	4	ltempf	0.4753	4	lbloodp	-1.0360
5	sga	0.5545	5	lserum	0.3539	5	seizure	0.5664
6	lserum	-0.5371	6	apgar5	0.3216	6	lurine	-0.2324
7	lbloodp	-0.5165	7	sga	-0.2704	7	sga	-0.2148
8	ltempf	-0.4617	8	po2fio2r	0.2443	8	ltempf	0.0671
9	seizure	0.4556	9	lurine	-0.0421	9	lserum	0.0607

In addition to the consideration of case distributions and the deviations of the statistical data which may make predictions inconsistent, we also performed tests to see whether the data size itself was a significant factor on our results. Five datasets were randomly

selected from the test set of the CNN database (recall that when the ANN tool was trained, one third of the data were used as a test set, and two thirds as a training set). Each set contained 59 cases. We compared the ANN tool's performance with similar data (in statistical values) but with different sample size. The responses of the ANN tool's output variables – Mort, LOS, Ventday (2-layer) and Ventday (3-layer) – were tested for each of the five datasets. Results are shown in tables from 5-12 to 5-15. In these tables, the “tool” column represents the classification performance of ANN tool (results from training the ANN); “EPIC” shows classification performance for EPIC database by ANN tool; “Test1” to “Test5” display classification performance for five datasets by the ANN tool.

Table 5-12 summarizes the results for “Mort” variable. In comparison with the sensitivity rate of the ANN tool (70.2%) obtained with the training set of data, the

Table 5- 12 Test results for variable “Mort”

	tool	EPIC	Test1	Test2	Test3	Test4	Test5
Sen (%)	70.2	25	40	50	75	80	43
Spec (%)	93	98	98	100	98	98	98
CR (%)	90.6	88.1	93.2	94.9	96.6	93.2	88.1
CP (%)	89.5	86.4	91.5	89.8	93.2	88.1	88.1

Table 5- 13 Test results for variable “LOS ”

	tool	EPIC	Test1	Test2	Test3	Test4	Test5
Sen (%)	71	63.2	72.7	67	64	85	55
Spec (%)	85.9	63.6	95.8	88	82	89	92
CR (%)	83.4	63.3	91.5	84.7	78	88	84.7
CP (%)	83.1	63.3	81.4	84.7	76.3	78	81.4

Table 5- 14 Test results for variable “Ventday” (2-layer)

	tool	EPIC	Test1	Test2	Test3	Test4	Test5
Sen (%)	64.3	40	67	60	56	69	50
Spec (%)	89	100	76	79.4	73.3	90	87
CR (%)	78.8	54.2	72.9	71.2	62.7	79.7	69.5
CP (%)	58.8	76.3	64.4	57.6	50.8	50.8	52.5

Table 5- 15 Test results for variable “ Ventday” (3-layer)

	tool	EPIC	Test1	Test2	Test3	Test4	Test5
Sen (%)	76.7	87	71.4	40	72.4	79	61
Spec (%)	90.6	0	89.4	91	93.3	87	87
CR (%)	84.8	66.1	83.1	69.5	84.7	83.1	74.6
CP (%)	58.8	76.3	64.4	57.6	50.8	50.8	52.5

corresponding sensitivity rates obtained from the five test data sets varied between -30% to +10% of the training data’s result. In regard to the specificity parameter (Spec %) the test results more closely matched the training results, with maximum variation of +7%.

Table 5-13 summarizes results for “LOS” variable. Here the sensitivity rates of the five test sets vary between -16% to +14% of the results for the training set whereas specificity rates vary between -4% to +10% of the training set. The corresponding results for the Ventday (2-layer) variable (Table 5-14) show sensitivity varies from -14% to +5% while specificity varies from -16% to +1%. Finally in Table 5-15, results for Ventday (3-layer), sensitivity rates vary from -37% to 2% while specificity rates vary between -4% to +3%. From the above results, we surmise that the fluctuation of prediction rates exist because the sample size is small. This is reasonable, because, in a sample that contains five positive cases, a single case’s difference in prediction will result in a 20% difference in the ANN’s computed sensitivity rate. However, some of the prediction rates for the EPIC database are far outside the resulting ranges of the five test sets. For example, the ANN tool’s results based on the EPIC data for Ventday (3-layer) yield a specificity of 0% for EPIC database (see Table 5-15), while the ANN tool’s corresponding rate is 90.6% for the CNN-based training data set. The influence of the small sample size of the EPIC database itself can not account for this significant difference. On the other hand,

comparison of the results for Ventday (2-layer) with Ventday (3-layer) in Table 5-14 and 5-15, the former is much better in its consistency when used with data from the EPIC database. This led us consider that the architecture of ANN may be one reason for the unstable prediction rate. Recall that ANNs learn by examples. In order to model the dataset closely, more input information should be added. However, a model that is too closely related to the training set is unable to generalize to accommodate new data. The ANN for Ventday (3-layer) was structured as a 3-layered network with five hidden nodes. It modeled the training data more closely than Ventday (2-layer), which means its generalization ability should be lower than that of the Ventday (2-layer) ANN. Since the data in the EPIC database is statistically different from that in the CNN database, the ANN predictions were correspondingly more divergent for the Ventday (3-layer) ANN, which modelled the CNN data more closely.

Based on above analysis, the bias of the prediction rates of the ANN tool may arise from several factors: The difference in case distributions of the two databases, the deviation of the statistic value of some input variables between the two datasets, the large difference of the sample size between the two data sets, as well as the differences in the architecture of the ANNs themselves. Also, the ANN training process may cause a bias. For example, when we trained the ANN using the CNN database, we divided the data into subsets for training and testing. Thus, not all the data was used for training. This would result in losing some information for ANN learning. To reduce this bias, we may resample, using a different split of data each time. This way, we build multiple neural networks. Then, we

average the predictions, and select the more useful result. However, this produces a cost in processing time.

All of these factors would – to a greater or lesser degree – contribute to the prediction fluctuations. Therefore, the ANN tool seems to be best applied to patients with similar characteristics found in the CNN database.

5.2.2 Results and Analysis for Physicians' Predictions

Table 5-16 shows the classification rates of the ANN tool and prediction rates for 59 cases in the EPIC database by physicians. By comparing the classification rates in the table, the Sensitivity, Specificity, and CCR of the ANN tool for the three outcome variables are better than those made by physicians, with only the exception that the sensitivity of Ventday (2-layer) is 10% lower than the sensitivity of the Ventday made by physicians. Thus, in general, the performance of the ANN tool was higher than that of physicians.

Table 5- 16 Classification rate of ANN tool (best results from testing ANN using CNN database) & classification rate for EPIC by physicians

	Mort		Ventday		LOS	
	Tool (CNN)	Physicians (EPIC)	Tool (CNN) (3-layer) (2-ayer)	Physicians (EPIC)	Tool (CNN)	Physicians (EPIC)
Sen (%)	70.2	50	76.7 63.4	73.3	71	63.2
Spec (%)	93	92	90.6 89	78.6	85.9	36.4
CCR (%)	90.6	86.4	84.8 78.8	74.6	83.4	53.3
CP (%)	89.5	86.4	58.8 58.5	76.3	83.1	63.3

Next, we compared the performance of the ANN tool and the physicians for the EPIC database. As we mentioned before, predictions of the ANN tool were affected by the differences of several factors between the CNN and EPIC database. Under this

circumstance, we also wanted to compare the tool with physicians to see if there were significant differences between them.

In Table 5-17, for three out of the six outcomes (negative outcomes) being predicted, the ANN tool made more correct predictions than physicians, for one outcome (positive outcome) the tool made equal correct predictions to physicians, and for two outcomes (positive outcomes) the tool made less correct predictions than physicians. In the “death” outcome, the total number of cases resulting in death was actually eight. The ANN tool predicted two of these deaths while the physicians predicted four. Since the total number

Table 5- 17 Real outcomes for 59 cases and cases correctly predicted by physicians and by ANN Tool

	Mortality		Ventday		LOS	
	Death (+1)	Survival (-1)	Ventday $\geq$ 24h (+1)	Ventday<24h (-1)	LOS>28 d (+1)	LOS $\leq$ 28 d (-1)
Real outcome	8	51	45	14	19	11
Physicians' outcome	4	47	33	11	12	4
Tool's outcome	2	50	18 (2-layer)	14(2-layer)	12	7

of the “death” cases were small, it is difficult to reach a conclusion that the physicians are sure doing better than the tool in predicting the death, as this may happened accidentally. For the Ventday outcomes, the real positive cases (Ventday $\geq$ 24h) are forty five. The physicians predicted thirty three cases correctly, while the tool predicted eighteen correctly. By consulting an expert at NICU in CHEO, we found that for the physicians, the Ventday predictions for patients in the EPIC database are relatively easy.

This is because the data of the fifty nine patients in the EPIC database were collected from the NICU at CHEO. All of these infants were transferred from other hospitals and, by the time they arrived, they were already assisted with artificial ventilation. The low sensitivity performance of the ANN tool with the EPIC database can be explained. An ANN trained on a data set with 900 good cases and 100 bad will bias its decision towards good cases, as this allows the algorithm to lower the overall error (which is much more heavily influenced by the good cases). If the representation of good and bad cases is different in the real population, the ANN's predictions may be wrong.[Neural Networks electronic textbook] In our ANN training, the negative cases were predominant in the CNN database. Then it was used in prediction on the EPIC database, which has a much higher proportion of positive cases than the CNN database.

For each case, to identify whether the predicted results of the physicians and of the tool were similar or not, would be helpful in analyzing whether the tool's predictions could be useful information for physicians. If the tool's predictions were always the same as the physicians', the tool would not be helpful for physicians. Table 5-18 shows the number of cases, out of fifty nine EPIC cases, for each output variable, where both physicians' and tool's predictions were wrong for a certain case, or physicians' predictions were correct while the tool's was wrong, or the physicians' predictions were wrong while the tool's was correct. The results show that for the positive results, the physicians' predictions were better than the tool's predictions, while for the negative outcomes, the current tool's correct predictions outnumbered the physicians'.

Table 5- 18 Comparison of physicians' and tool's prediction results for each case

Physicians' and Tool's prediction		Mort	Ventday	LOS
Positive results (+1)	Both wrong	4	8	4
	Physicians right & Tool wrong	2	19	3
	Physicians wrong & Tool right	0	4	3
Negative results (-1)	Both wrong	0	0	4
	Physicians right & Tool wrong	1	0	0
	Physicians wrong & Tool right	2	3	3

5.3 Discussion

This was a pilot study in which the physicians' attitude toward using the ANN tool was measured, and the performances of the ANN tool and of the physicians was compared in predicting outcomes. The results are encouraging. Most of the physicians who participated in the study showed positive attitudes to the ANN tool, even though they have not used it. This indicates that physicians need more useful information to help them with their decision-making. From the study, we found that physicians believed that the ANN tool would be helpful in two aspects: one is the estimate of outcomes; the other is the counselling of parents. The prediction results of the ANN tool could help physicians' confidence when the results of the tool are similar to the results made by physicians.

5.3.1 Strengths and Limitations of the ANN Tool in Making Predictions

Making a medical diagnosis and prognosis are complex procedures. Physicians require large amounts of information to make decisions. One of the advantages of the ANN tool is that this tool allows a physician to access a huge quantity of patient data, to analyze that data by using computer techniques, and glean valuable information with which to make medical decisions. For many medical databases, relationships within the data are complex and nonlinear, as can be seen in the CNN database. For example, the best results for the outcome variable “Ventday” – that is, the best match to the database’s outcomes -- were obtained by training the ANN with a 3-layered structure. This would indicate that the relationships between monitored parameters and the number of required ventilation days (Ventday) are nonlinear. This ability to model non-linear relationships between several variables is one reason that ANNs are widely used in medical applications.

Another advantage of the ANN tool is that it is easily updated. This characteristic is particularly important in medical applications, as dramatic changes in the underlying data could occur at anytime due to improvements in the diagnosis procedures or treatment protocols, which may significantly alter the outcomes. For example, a new drug or medical device may be capable of saving many patients’ lives. The new results caused by such changes can be added to the training data set, thereby allowing the ANN to do the re-training; as a result, new data characteristics are learned by ANN and the outcome predictions are changed accordingly. It should be pointed out that, when the training data is updated, it is not enough to simply add the new data. Data that is out of date should be deleted as well.

Although the ANN tool does have the strength to help physicians with their decision-making, there still exist some limitations for its application. During the evaluation process, we found that the prediction rates of the tool were uncertain when the cases in the sample were not similar to those in the training set. And even when the cases in the samples were similar to the training set, there were fluctuations. Therefore, the question was raised of how we might gauge the accuracy of the ANN tool's prediction. It would be helpful if an estimate of the ANN prediction's accuracy was computed for individual cases. Such a model/tool would not only give an output value, but an indication of the accuracy of that output as well. Currently, ANNs are often viewed as black boxes: they can model a variety of relationships, but do not provide what these are. Many researchers are trying to overcome some of these limitations. For example, studies on quantifying the uncertainty in measurement of ANN by confidence interval (CI) are in progress [Arslan O. *et al.* 2004, Wansink G. *et al.* 2001, Yang L. *et al.* 2000]. A CI gives an estimated range of values which is likely to include an unknown population parameter, the estimated range being calculated from a given set of sample data. For example, if the prediction value of death of an infant is 0.8 by neural network, the relative risk of death of an infant may be expressed as 0.8 (95% CI, 0.6 to 1.0), meaning that the relative risk interval between the value 0.6 and 1.0 has a 95% probability of containing the true death. Wider intervals indicate lower precision; narrow intervals indicate greater precision.

Another limitation is that even though the ANN tool is flexible and can be updated, it can not respond to changes so quickly. There is always a delay for updating the tool because of the need to accumulate new cases and adding them to databases.

Also, the performance of the ANN tool is affected by the deviation of the data, which means that the results from the ANN training can be generalized to the population similar to the ANN training datasets. It should be recognized that the results with ANNs always depend on the data with which they were trained. An ANN can only learn from cases that are presented to it. We cannot expect the tool to make a correct prediction when it encounters one of the previously-unseen cases. Therefore, the ANN tool would best be used for cases that are similar to the population used in the ANN training. Although this shortcoming could be remedied by enlarging the training data, the enlargement is not unlimited.

5.3.2 Strengths and Limitations of Physicians in Making Predictions

Diagnosis and prognosis are very important in medicine. Before physicians make a diagnosis, they need to collect large amounts of information, such as anamnesis, symptoms, physical examination, laboratory and instrument data, clinical data, etc. As the human body is very complicated and exhibits individual variations, diagnosis should be made on a case by case basis. One of the strengths of physicians is that their decision making capacity is flexible -- they are able to assess patients under a wide range of circumstances. In addition, they are quick to respond to changes. Experience is especially important in medicine. Physicians obtain experience through daily clinical practice. The number of patients treated by a physician is necessarily limited by time, location, demographics, budgets, and capacity. Modern technologies make it possible to enrich physicians' experience by providing a huge amount of information through medical databases, which accumulate others' experience.

Chapter 6 Conclusions, Contributions and Future Work

6.1 Conclusions

This study was meant to evaluate the ANN tool's ability to help physicians in their clinical work. Two pilot studies, one attitude test and one laboratory experiment, were performed to measure the physicians' attitude toward using the ANN tool, and to test the performance of the ANN tool when compared with that of physicians. The results of the study found that physicians' responses to the ANN tool were fairly positive. Most of them believe that the tool can help them in estimating the mortality, the patient's length of stay in NICU and the likely length of time a ventilator would be needed. In addition, they believe that the ANN tool will be useful for counseling parents and making decisions.

As the tool is new to physicians and they have little experience using it, a few physicians were concerned about its ability to predict and the possibility of misuse. This reminds us that before the ANN tool is formally applied to the clinical world, it is important to train physicians on its reliability (or lack thereof) of the ANN tool's performance and on how to use it. The tool's results should neither be regarded as a "prescribed course of action" in regard to a patient, nor should they be given to parents alone as "raw data" without accompanying professional counseling. In short, the ANN tool should never replace a physician's decision-making responsibility. Instead, as one additional kind of information among the many kinds of information needed, it is merely intended to assist physicians with their decision-making.

The results of the experiments demonstrate that the testing results from the ANN training using the CNN database are higher than the performance of the physicians on the EPIC database in terms of sensitivity, the specificity and the CCR. When used with the EPIC database – in which the population is different from the CNN database – the ANN tool also showed somewhat comparable performances with those of physicians.

6.2 Contributions

The following is the summary of the contributions to new knowledge of this thesis.

1. Developed an ethic proposal for this study, and was submitted and approved by the Ethics Review Committee at CHEO.
2. Developed and applied a questionnaire and approach to assess the attitudes of physicians on using the ANN tool;
3. Created scenarios in order to make it possible for physicians doing prediction based on the data in the database.
4. Identified the prediction strengths and limitations of the tool and of the physicians so that MIRG can now concentrate on improving the tool to the point of clinical applicability.
5. Identified a key variable not in the minimum data set of the SNAPPEII: “gestational age”;

6.3 Future Work

The following are recommended areas for future study with a view to improving the ANN tool.

1. Training the ANN by adding the variable ‘gestational age’, which was found important to physicians, to see if it will improve the ANN results.
2. Using an appropriate statistical method to define the confidence intervals for predicted outputs.
3. Researching the effect of input data deviations on the ANN’s predicted outcomes.
4. Conducting further pilot studies at CHEO on refinements done of the tool and upgrades.
5. Conducting a clinical trial in multi-center NICUs before the ANN tool goes into clinical practice.

References

Adelman L and Riedel SL: Handbook for evaluating knowledge-based systems. Kluwer Academic Publishers. 1997.

Alberts B, Bray D, Johnson A, Lewis J, Raff M, Roberts K, and Walter P: Essential Cell Biology — An Introduction to the Molecular Biology of the Cell. Garland Science Publishing, New York, 1998.

American Institutes For Research: Usability Engineering.
<http://www.measuringusability.com/about.html>

Anderson D, McNeill G.: Artificial Neural Networks Technology—A DACS State-of-the-Art Report. Contract Number F30602-89-C-0082 (Data & Analysis Center for Software) ELIN: A011, August 20, 1992, Prepared for: Rome Laboratory, RL/C3C

Arslan O. et al A Practical Method of Confidence Estimation for Neural Networks in Image Classification IJSIT Lecture Note of International Conference on Intelligent Knowledge Systems, Vol.1, August 2004

Atienza F, et al.: Risk Stratification in Heart Failure Using Artificial Neural Networks. Proc. AMIA Annual Fall Symposium 2000. pp. 32-36. Available:
<http://www.emedicine.com/ped/topic2767.htm>

Bartlett PL.: For valid generalization, the size of the weights is more important than the size of the network. In Mozer, M.C., Jordan, M.I., and Petsche, T., (eds.) Advances in Neural Information Processing Systems 9, Cambridge, MA: The MIT Press, 1997; pp.134-140.

Baxt W.G.: Use of an artificial neural networks for data analysis in clinical decision-making: the diagnosis of acute coronary occlusion. Neural Comp 1990; 2(4):480-489.

Baxt W.G.: Use of an artificial neural networks for diagnosis of myocardial infarction. Ann Intern Med 1991;115:843-848.

Buchman TG, et al.: A comparison of statistical and connectionists models for prediction of chronicity in a surgical intensive care unit. Crit Care Med 1994;22(5):750-762.

Buskard T, Stevenson M, Frize M, Solven FG.: Estimation of ventilation, length of stay, and mortality using artificial neural networks. Proc.Can Conf Elec Comput Eng 1994:726-729.

Carter BS.: Ethical Issues in Neonatal Care, E-Medicine.com,2004.

Clarke K et al.: A Methodology for evaluation of knowledge-based systems in medicine. *Artif Intell Med.* 1994 April; 6(2):107-121

Dawson AE, Austin RE Jr, Weinberg DS: Nuclear grading of breast carcinoma by image analysis. Classification by multivariate and neural network analysis. *Am J Clin Pathol*, 1991; 95(4Suppl1): S29-37.

Dayhoff Judith E, DeLeo James M.: Artificial neural Networks—Opening the Black Box. American Cancer Society, Cancer Supplement, 2001 April; 91(8):1615-35.

DeFigueiredo RJ, Shankle WR, Maccato A, Dick MB, Mundkur P, Mena I, Cotman CW.: Neural-network-based classification of cognitively normal, demented, Alzheimer disease and vascular dementia from single photon emission with computed tomography image data from brain. *Proc Natl Acad Sci USA*, 1995; 92(12): 5530-4

Ennett CM, Frize M, Charette E.: Improvement and automation of artificial neural networks to estimate medical outcomes. *Med Eng Phys.* 2004 May; 26(4):321-8.

Ennett CM, Frize M. Scales N.: Evaluation of the logarithmic-sensitivity index as a neural network stopping criterion for rare outcomes. *Proc IEEE ITAB* 2003.

Ennett CM, Frize M, Scales N.: Logarithmic-sensitivity index as a stopping criterion for neural networks. *Proc IEEE EMBS-BMES* 2002.

Ennett CM, Frize M, Walker CR.: Influence of missing values on artificial neural network performance. *Medinfo* 2001; 10(Pt 1):444-53

Ennett CM, Frize M, Charette E.: Automated Optimisation of Neural Networks Performance. *MEP* 2004; 26:321-328.

Ennett CM.: Imputation of Missing Values by Integrating Artificial Neural Networks and Case-based Reasoning. PhD thesis, System and Computer Engineering, Carleton University, Ottawa, ON, Canada. 2003

Ennett M, Frize M, Scales N.: Logarithmic-Sensitivity Index as a Stopping Criterion for Neural Networks. *IEEE/EMBS*, Fall 2002, Texas.

Ennett CM, and Frize M.: An Investigation into Strengths and Limitations of Artificial Neural Networks: An Application to an Adult ICU Patient Database. Published as a poster paper, *Proceedings of A.M.I.A.(American Medical Informatics Association) Annual Symposium*, Florida, November 1998; 7-11: 998.

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

Fogel DB, Wasson EC 3rd, Boughton EM: Evolving neural networks for detecting breast cancer. *Cancer Lett*, 1995; 96(1): 49-53.

Frize M, Buskard T, Stevenson M, Solven FG Dunfield V.: Preliminary results of the development of an artificial neural network for the intensive care. Proc. IEEE-EMBS 1993; 296.

Frize M, Solven FG, Stevenson M, Nickerson B, Buskard T, Taylor K.: Computer-assisted decision-support systems for patient management in an intensive care unit. Medinfo 1995; 8 Pt. 2:1009-1012.

Frize M, Solven FG, Stevenson M, Nickerson B, McGowan HCE.: Information technologies approach and development for various medical applications. Proc. Can Conf. Elec Comput Eng 1996; 365-368.

Frize M, Ennett CM and Charette E.: Automated Optimization of the Performance of Artificial Neural Networks to Estimate Medical outcomes. Proc. 3rd ITAB Conference (Information Technology Applications in Biomedicine) and ITIS (International Telemedical Information Society), Washington, November 2003; 168-173.

Frize M, Ennett C. M, Stevenson M. and HCE Trigg: Clinical Decision-Support Systems for Intensive Care Units Using Artificial Neural Networks. Medical Engineering & Physics. 2001; 23:217-225.

Frize M, Walker RC, Ennett CM.: Development of an Evidence-Based Ethical Decision-Making Tool for Neonatal Intensive Care Medicine. 25th Silver Anniversary International Conference of the IEEE Engineering in Medicine and Biology Society. Cancun, Mexico, September 2003

Frize M, Yang L, Walker RC, and O'Connor A.: Conceptual Framework of Knowledge Management for Ethical Decision-Making Support in Neonatal Intensive Care. Transactions of Information Technology in Biomedicine. 2005.

Geman S, Bienenstock E, and Doursat R.: Neural Networks and the Bias/Variance Dilemma. Neural Computation, 1992; 4, 1-58.

Greenes RA, Tarabar DB, Krauss M, Anderson G, Wolnik WJ, Cope L, Slosser E, McGowan HCE, Stevenson M and Frize M.: The Need for Standardized Reporting of Medical Applications of Artificial Neural Networks. CMBEC, Charlottetown, June 1996; 26-29.

Hersh W: Knowledge management as a decision-support method. Computers in Biomedical Research 1989; 22(2), 113-135.

Jeff Sauro: Measuring usability. <http://www.measuringusability.com/about.html>

Johnston ME, Langton KB, Haynes RB and Mathieu A.: Effects of computer-based clinical decision support systems on clinician performance and patient outcome. Annals of Internal Medicine. 1994; 120(2), 135-142.

Kimball JW, Kimball's Biology Pages:

<http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/N/Neurons.html>

Knaus WA, Zimmerman JE, Wagner DP, Draper EA, Lawrence DE.: APACHE – acute physiology and chronic health evaluation: a physiologically based classification system. Crit Care Med. 1981; 9(8):591-7.

Kornel Papik et al.: Application of neural networks in medicine — a review, Med Sci Monit, 1998; 4(3): 538-546.

Kuntz G.R and Hammond J.: Predicting outcomes after closed-head injury (CHI) using a cascade-correlation artificial neural network. Crit Care Med 1994; 22(1) A25.

Leveson N, Clark S. T.: An Investigation of the Therac-25 Accidents. IEEE Computer, July 1993; Vol. 26, No. 7, pp.18-41.

Liu H.: Pilot Studied of Physician-Computer Interaction Using the IDEAS Software For Neonatal ICUS. Master's thesis, 1998 System and Computer Engineering, Carleton University, Ottawa, ON, Canada. 2003

Livingstone DJ, Manallack DT, Tetko IV.: Data modeling with neural networks: Advantages and limitations. Computer-Aided Molecular Design 1997; 11:135-142.

Maclin PS, Dempsey J.: Using artificial neural networks to diagnose hepatic masses. J Med Syst, 1992; 16(5): 215-25.

McCullogh W.S, and Pitts W.H.: A logical calculus of the ideas immanent in nervous activity, Bulletin of mathematical biophysics. 1943; vol. 5, p.115-133.

Mehrotra K. et al.: Elements of artificial neural networks. Cambridge, MA; A Bradford Book, MIT Press, 1997.

Miller RA, Pople HE. and Myers JD.: An experimental computer based diagnostic consultant for general internal medicine. New Eng. J. Med. 1982; 307,468-476.

Nazeran H, Rice F, Moran W, Skinner J.: Biomedical image processing in pathology: a review. Australas Phys Eng Sci Med, 1995; 18(1): 26-38.

Neural Networks electronic textbook, <http://www.statsoft.com/textbook/stneunet.html>
© Copyright StatSoft, Inc., 1984-2003, STATISTICA is a trademark of StatSoft, Inc.

Nielsen J : Jakob Nielsen's Alertbox, March 19, 2000: Why you only need to test with 5 users.

Nohr C: The evaluation of expert diagnostic systems. How to assess outcomes and quality parameters. *Artificial Intelligence in Medicine* 1994; 6(20), 123-135.

Oppenheim AN.: Questionnaire Design and Attitude Measurement. London, Heinemann, 1982.

Pearl J.: probabilistic Reasoning in Intelligent System. San Mateo, California: Morsa Kaufaman, 1988.

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

Pesonen E, Eskelinen M, Juhola M.: Treatment of missing data values in a neural network based decision support system for acute abdominal pain. *Art Intell Med.* 1998; 13, 139-46.

Ravdin PM, Clark GM, and Hilsenbeck SG.: Demonstration that breast cancer recurrence can be predicted by neural network analysis. *Breast Cancer Res Treat*, 1992; 21(1):47-53.

Reisman Y.: Computer-based clinical decision aids. A review of methods and assessment of systems. *Medical Informatics* 1996; 21(3) 179-197.

Richardson DK *et al.*: Score for Neonatal Acute Physiology–A Physiologic Severity Index for Neonatal Intensive Care. *Pediatrics* 1993; 91(3):617-623

Richardson DK, Corcoran JD, Escobar GJ and Lee SK.: SNAP-II and SNAPPE-II: Simplified newborn illness severity and mortality risk scores. *The Journal of Pediatrics.* 2001; 138, 92-100.

ROC online 1: <http://gim.unmc.edu/dxtests/roc3.htm>

ROC online 2: <http://www.cmh.edu/stats/ask/roc.asp>

Scales N.: Mortality prediction in an ICU and coronary surgery database using neural networks. Term project, Department of Systems and Computer Engineering, Carleton University, Dec 2001.

Schwartz D, and Lellouch J: Explanatory and pragmatic attitudes in therapeutic trials. *Journal of Chronic Diseases* 1967; 20,637-648.

SGA Tool online: <http://www.sfar.org/scores2/snap22.html#help>

Snowden S, Brownlee K.G, and Dear P.R.F.: An advisory system for artificial ventilation of the newborn utilizing a neural network. *Med Inform (Lond).* 1993 Oct-Dec; 18(4):367-76.

Snowden S, Brownlee K.G, and Dear P.R.F.: An expert system to assist neonatal intensive care. J Med Eng Technol 1997 Mar-Apr; 21(2):67-73, Vol. 21, No. 2, Pages 67-73.

Tong Y, Frize M, and Walker R.: Extending Successful Predictions of Ventilation Needs using Artificial Neural Networks from Adult to Neonatal ICU Patients IEEE Transaction on Information Technology in Biomedicine. June 2002; 6(2), 188-191

Tong Y, Frize M, Walker R.: Estimating ventilation using artificial neural networks in intensive care units. First Joint BMES/EMBS (IEEE) Conference, Atlanta, GA, Oct.1999

Tourassi GD, Frederick ED, Floyd CE Jr and Coleman RE.: Multifractal texture analysis of perfusion lung scans as a potential diagnostic tool for acute pulmonary embolism. Comput Biol Med 2001; 31:15.25.

Tourassi GD, Frederick ED, Vittitoe NF and Coleman RE.: Fractal texture analysis of perfusion lung scans. Comput Biomed Res 2000; 33:161.171.

Tu JV, Guerriere MRJ.: User of a neural network as predictive instrument for length of stay in the intensive care unit following cardiac surgery. Comput Biomed Res 1993; 26:220-229.

Walczak S. and Nowack W.: An Artificial Neural Network Approach to Diagnosing Epilepsy Using Lateralized Bursts of Theta EEGs. Med Systems 2001; 25 (1): 9-22.

Walker CR, Frize M, Tong Y.: Use of an Artificial Neural Network (ANN) to Analyse Duration of Ventilation in Neonatal Intensive Care Unit (NICU) Patients. Pediatric Academic Societies (APS/SPR) and American Academy of Pediatrics Joint Meeting, Boston, MA, 2000 May; 12-16.

Wansink G. *et al.*: A new Confidence Bound estimation method for neural networks, an application example. EAGE 63rd Conference & Technical Exhibition — Amsterdam, The Netherlands, 11 - 15 June 2001

Wasson JH, Sox HC, Neff RK and Goldman L.: Clinical prediction rules: applications and methodological standards. New England Journal of Medicine 1985; 313, 793-799.

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.

Werbos P.: Beyond regression—new tools for prediction and analysis in the behavioral sciences. PhD Thesis, Harvard University, Cambridge, MA. August 1974

Wyatt J, Spiegelhalter D.: Evaluating medical expert systems: what to test and how? Med Inform (London). 1990 Jul-Sep; 15(3):205-217.

Yang L, Kavli T, Carlin M., Clausen S, and de Groot P. F. M.: An Evaluation of Confidence Bound Estimation Methods for Neural Networks, ESIT 2000: 322 - 329.

Yang L, Frize M, Walker RC and Catley C.: Towards Ethical Decision Support and Knowledge Management in Neonatal Care. Proc. IEEE EMBS Conference, SF, Sept. 2004

Appendix 1 SNAP Variable List and Scoring System [Richardson *et al.* 1993a]

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
pO ₂	Low	>65	50-65	30-50	<30
pO ₂ /FiO ₂ ratio	Low	>3.5	2.5-3.5	0.3-2.49	<0.3
pCO ₂	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 ratio	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	-	-	-	-
- bili for birth weight >2kg		<15	15-20	>20	-
- bili/kg 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 (<5 th percentile)	5 points

Appendix 2 Questionnaire for Pilot Study 1

Description:

Our research group is now developing a decision-support system for potential future clinical application. ANN tool is one of the products of this system which uses artificial neural networks to estimate the outcomes for the single newly admitted patient to NICU: *mortality*, the *duration of artificial ventilation*, and the *length of stay* in the NICU. This questionnaire is designed to collect information on the attitude of physicians towards using this tool and suggestions on the improvement of the tool.

In order to analyze the results more objectively, before starting the questionnaire, please answer a small amount of demographic information about you:

1. Where (in which country/ies) did you receive your medical education?

Answer:

2. How many years have you worked as a neonatologist?

Answer:

.

3. Have you worked in any medical field other than neonatology? If yes, what field/s, and for how many years?

Answer:

Questions	Please select one for each answer 1 very likely—3 uncertain—5 very unlikely				
	1	2	3	4	5
1. Would you use the ANN tool to assist your diagnosis and treatment of patients if the tool's outcome predictions were similar to your prediction as a physician?					
2. Would you use the ANN tool to assist your diagnosis and treatment of patients if the tool's outcome predictions were significantly different from your prediction ?					
3. If your estimation of probability of death for a patient were significantly different from the ANN tool's estimation, how likely would you be to accept each prediction (below)?					
Likely to accept your estimate					
Likely to accept ANN's estimate					
4. Do you think the information (ANN tool's outcome prediction) would help you in estimating mortality in terms of short-term survival (until discharge)?					
5. Do you think the information would help you in estimating the likely length of time a ventilator would be needed?					
6. Do you think the information would help you in estimating the patient's length of stay in NICU?					
7. Do you think the information would help you in estimating the risk of complications (e.g., IVH, PVL) in the patient?					
8. Do you think the ANN tool would speed up your decision-making process?					

9. The system uses specific **input** variables to predict outcome (see list below.)
Please rank from 1-5 the five most important variables you would use to estimate probable outcome for a newborn infant (feel free to add variables to this list for your ranking of the top 5.)

<u>Variable</u>	<u>Description</u>	<u>[Ranking]</u> (1 most important)
Bthwt	birth weight	<input type="text"/>
Sga	whether small for gestational age	<input type="text"/>
apgar5	Apgar score at 5 minutes	<input type="text"/>
Hbloodp	highest blood pressure in 1 st 12 hours	<input type="text"/>
Lbloodp	lowest blood pressure in 1 st 12 hours	<input type="text"/>
Ltempf	lowest temperature in 1 st 12 hours	<input type="text"/>
Hrespr	highest respiratory rate in 1 st 12 hours	<input type="text"/>
Lserum	lowest serum pH in 1 st 12 hours	<input type="text"/>
Po2fio2r	lowest pO ₂ :FiO ₂ ratio in 1 st 12 hours	<input type="text"/>
Hpco2	highest pCO ₂ in 1 st 12 hours	<input type="text"/>
Hsodium	highest serum Na in 1 st 12 hours	<input type="text"/>
Lgluc	lowest blood glucose in 1 st 12 hours	<input type="text"/>
Lplt	lowest platelet count in 1 st 12 hours	<input type="text"/>
Lurine	urine output in 1 st 12 hours	<input type="text"/>
Seizure	whether multiple seizures	<input type="text"/>
Other (please describe)	<input type="text"/>	<input type="text"/>
Other (please describe)	<input type="text"/>	<input type="text"/>
Other (please describe)	<input type="text"/>	<input type="text"/>
Other (please describe)	<input type="text"/>	<input type="text"/>

10. Other than the variable list with those you may have added for your ranking now included in 10 above, are there other patient data that you think are important for making your diagnosis, deciding treatment/management of the patient or predicting prognosis?
-

11. Other comments about the ANN tool.

(E.g., would **outcomes** other than those in Q.4-7 be useful for clinical decision-making? In your opinion, what are the most positive **benefits** of the system? Do you see the possibility of anyone **misusing** the system? Any other comments?)

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

Appendix 3 The Specifications for SNAP [Richardson DK. *et al.* 1993]

1. rely on physiology rather than diagnosis or therapy to reduce the impact of case mix, diagnostic labeling, changes in therapy over time, or differences in styles of practice between NICUs;
2. provide information beyond that available from traditional risk factors such as birth weight, gestational age, sex, race, and Apgar score;
3. have established validity and reliability in a variety of practice settings;
4. be clinically plausible to facilitate its use by physicians as well as health service researchers;
5. make use of readily available clinical information to avoid the need for special tests and to facilitate its adaptation to retrospective review;
6. have sufficient range to distinguish among the broad spectrum of cases seen in NICUs;
7. reflect status on admission in order to separate the contributions of obstetric and demographic risks from the effectiveness of NICU treatment.

Appendix 4 Proposal for Project

Proposal for Project

For Ethics Review at CHEO

**PROJECT: COMPARISON OF PREDICTIONS BY AN ARTIFICIAL NEURAL
NETWORKS TOOL AND PHYSICIANS OF OUTCOMES OF NEONATAL
INTENSIVE CARE UNIT CASES**

Medical IDEAS Research Group

Department of Systems and Computer Engineering

Carleton University

Department of Paediatrics (Division of Neonatology)

CHEO

January, 2005

The CHEO Research Ethics Review Board has previously approved a number of projects of MIRG, including our CFI-funded NICU Data Collection Infrastructure and the first phase of our PSI-funded Ethical Decision Support Tool study. Our Artificial Neural Network (ANN) and Case-Based Reasoner (CBR) tools have been well studied for their accuracy in projecting several NICU outcomes (including survival, ventilation duration, etc.) on patients derived from our own databases and that of the Canadian Neonatal Network (full background and references were included in our PSI grant proposal.)

However, the ANN and CBR systems developed by MIRG are still currently in the lab experiment phase (as you know, we are only slowly moving into the clinical domain through studies such as the ethical decision support tool study.). Now, graduate student Lili Qi wishes to do a small side-study to compare the accuracy of the tools with that of experienced neonatal physicians in projecting these outcomes.

The long term objective of MIRG's (Medical Information technology Research Group) research team is to use artificial intelligence approaches to create a number of promising clinical decision-support systems which will ultimately be integrated into one system to support decision-making in the NICU. The value of such systems however depends on their ability to improve on current decision-making. We have presented our work to the members of the Division of Neonatology at CHEO and those physicians support the concept but would like to participate in a small study to assess whether the systems actually do improve on their own predictive abilities! (If in fact physicians already project patient outcomes with accuracy approaching that of the systems, the new systems would not necessarily be of value.)

The objective of Ms. Qi's project is to evaluate the ANN tools for their usefulness in NICU by measuring the accuracy of their estimations of the likelihood of infants' surviving, the estimated duration of artificial ventilation, and length of stay (LOS) in comparison with estimations by physicians in the NICU at CHEO. The 'cases' will be derived from real data in the CHEO subset of data collected for the Canadian Neonatal Network EPIC study; however, these data contain no identifying information of any kind.

ANN model estimates and physician estimates will be compared with the actual outcomes of the original cases. Information on the physicians' opinions on the ANN systems will also be collected by conducting a questionnaire (attached separately.) In order to analyze the results more objectively, a small amount of demographic information on participating physicians will be collected*. Informed consent will be obtained from each physician before his/her participation in the study (proposed consent form attached separately.)

*We have decided that only three questions are necessary for this:

1. Where (in which country/ies) did you receive your medical education?
2. How many years have you worked as a neonatologist?
3. Have you worked in any medical field other than neonatology? If yes, what field/s, and for how many years?

Appendix 5 Consent Form for Participating the Project

Development of an evidence-based ethical decision-support tool for physicians and parents in the Neonatal Intensive Care Unit

Explanation of the Study and Health Professional Consent

C Robin Walker, MB, ChB, FRCPC, University of Ottawa, Ottawa, ON 613-737-2393 X 2415
Monique Frize, P Eng, OC, Carleton University, Ottawa, ON (613) 520-2600 X8229
Lili Qi, M.Sc, Carleton University, Ottawa, ON (613) 520-2600 X5586

SPONSOR: Physicians Services Incorporated Foundation (PSIF)

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, risks and precautions 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.

Procedures

You will be asked to predict the outcomes of a series of cases using data that will also be supplied to the 'intelligent' system. You will not receive any information that could allow you to identify real cases, although the data have been derived from actual cases contained in a database used for our research. Your predictions and those of the ANN 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.

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 of evidence 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 Robin Walker at 737-7600 X2415. 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 X3624.

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 a 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