Development of Models for Predicting Severity of Childhood Falls

by

Qi Li
(Master of Science, Southeast University, China, 1999)
(Bachelor of Medicine, China Medical University, China, 1995)

A thesis submitted to the
Faculty of Graduate Studies and Research

In partial fulfillment of the requirements for the degree of

Master of Applied Science in Biomedical Engineering

Ottawa-Carleton Institute for Biomedical Engineering (OCIBME)

Department of Systems and Computer Engineering

Carleton University

Ottawa, Ontario, Canada, K1S 5B6

September 2010

Your file Votre référence
ISBN: 978-0-494-71516-1
Our file Notre référence
ISBN: 978-0-494-71516-1

Abstract

This work analyzed data collected by the Canadian Hospitals Injury Reporting and Prevention Program (CHIRPP) in 2007 to conduct research on childhood injuries resulting from falls. Three models were developed to predict the severity of childhood fall injuries - by using logistic regression, decision trees, and artificial neural networks. The data were collected upon arrival to the emergency room after the children's fall. Due to the large number of variables included in this dataset, logistic regression analysis was used to identify significant predictors which were then used as inputs for decision trees and artificial neural networks. Although all three models showed very good predictive ability regarding this research issue, we concluded that decision trees based on See5 was the most convenient and efficient approach to predict the severity of childhood fall injuries using CHIRPP data as much less work was needed in data preparation compared with the other two methods.

Acknowledgements

I would like to thank my two supervisors, Dr. Monique Frize and Dr. James Green for their constant encouragement in my thesis research work. Dr. James Green guided through my entire graduate study at Carleton University. Dr. Monique Frize introduced me to some important contacts and to artificial neural networks which was most helpful in the completion of this thesis. I am sincerely grateful for the opportunity to be part of the MIRG (Medical Information-technologies Research Group), which is led by Dr. Frize, to develop my knowledge and skills in this exciting field.

I would like to thank Jeff Gilchrist and Nicole Yu for their experience and great help on the neural network techniques.

I appreciated the useful information about the CHIRPP dataset provided by Steven McFaull and Corrine Langill. They also provided me valuable ideas for this research through discussion.

Table of Contents

List of Tables

LIST OF FIGURES

Nomenclature

ANN Artificial Neural Network

ASE Asymptotic Standard Error

AUC Area Under ROC Curve

CART Classification and Regression Trees

CCR Correct Classification Rate

CHIRPP Canadian Hospitals Injury Reporting and Prevention Program

CYIR Child and Youth Injury in Review

DLBP Double Layer Back Propagation

DT Decision Tree

GLM Generalized Linear Model

LR Logistic Regression

MIRG Medical Information Technologies Research Group

MLE Maximum Likelihood Estimation

MLP Multilayer Perceptron

MSE Mean Squared Error

NICU Neonatal Intensive Care Unit

PHAC Public Health Agency of Canada

ROC Receiver Operating Characteristic

CHAPTER 1. INTRODUCTION

1.1 Motivation

1.1.1 Status of Childhood Falls in Canada

According to the World Health Organization's definition, a fall is "an event which results in a person coming to rest inadvertently on the ground or floor or other lower level" [WHO online]. While falling is a normal part of childhood development and learning experience, it is also an important cause of unintentional injuries for children.

In Canada, unintentional injuries are the leading cause of mortality, morbidity and disability for children and youth [PHAC online]. Unintentional falls are also the most common types of childhood injury seen in emergency rooms [Flavin MP et al 2006] and the leading cause of injury-related hospitalization for children of 0-14 years old resulting in high social and economic costs. Boys are overrepresented in both fatal and non-fatal fall related injuries [CYIR 2009].

The leading locations of falls for children are playground, schools, and residential yards. For children younger than one year old, falls from furniture and car seats are the most common types. Older children usually fall from stairs, from a play structure or by being pushed [Crawley T 1996].

Falls are not only an important issue for children. Falls are also the most common types of injuries for the Canadian population in general. It was the most costly type of injury according to both the overall and direct health care costs;

approximately $4.5 billion or 42 percent of direct costs of injuries were due to falls in Canada in 2004 [SMARTRISK. 2009]. Furthermore, among all types of injuries, falls are the leading cause for permanent disability, which may require treatment for the patient's entire life. Approximately $1.7 billion was spent in 2004 to treat patients with permanent disabilities resulting from falls in Canada. Therefore, fall injuries cost Canadians about $6.2 billion in 2004, which accounts for 31 percent of total expenses for all injury types [SMARTRISK. 2009].

The Government of Canada has recognized that the prevention of fall-related injuries among children is of paramount importance and a national prevention strategy is required [Public Health agency of Canada online]. A leading strategy to prevent injuries from falls in many high-income countries is to regularly replace or modify unsafe products, such as nursery furniture, playground equipment, sport and recreational equipment [Mackey M et al. 2006].

1.1.2 Severity of Falls

Most pediatric falls are of little consequence or only of modest severity; however, some falls are severe and may result in permanent disability or even death. The severity of fall injuries depends on a number of circumstance conditions, such as kinetic energy transition, body position, and impact surface. Generally, greater height from which a child falls causes more severe injury [Committee on Injury and Poison Prevention 2001].

Different types of falls may cause injuries at different body parts, such as head injuries, spine injuries, orthopedic injuries, thoracic injuries, and abdominal injuries [Wang MY et al. 2001]. Injuries at different body parts will lead to various

consequences. Head injuries are the most common reason for a child to be admitted to hospital after a fall, especially in young children. Sometimes a fall from several feet could even produce cranial vault fractures for infants, which carries a significant risk of death and long-term consequences [Pickett W et al. 2003].

A Canadian study indicated that 36% of infants younger than one year admitted to an emergency department following a fall had significant head injuries. Falls were responsible for 90% of all head injuries seen in the emergency department [Pickett W. et al, 2003]. Falls are also the most common cause of fatal and severe head injuries among children in other developed countries [Williamson LM et al. 2002]. Occult visceral and intracranial injuries are also dangerous outcomes from falls. Since these injuries are often difficult to detect, these are two of the main causes of death from falls in children [Wang MY et al. 2001].

1.1.3 Problem Statement

Although precautions and safety equipment can decrease the severity of falls, effective and immediate emergency response and trauma care are very important to save lives after the injury has occurred. For those cases with severe fall injuries, either staying in the emergency room for observation or being admitted to hospital immediately to receive complete treatment can help them to avoid long-term consequences, or even to save their lives.

On-the-spot treatment in emergency rooms with necessary follow-up may be sufficient for some modest fall-related injuries. Minor injuries may not require any treatment at all. However, necessary medical advice is helpful for most patients with

a fall-related injury to effectively avoid repeat injuries in the future. Those patients with modest or minor injury can be discharged immediately. The severity of a fall will determine the treatment strategy. Therefore, the question arises on how to correctly and immediately differentiate patients with life- or limb- threatening injuries from those with less severe injuries, so that the severe patients can be treated in a timely and proper fashion. This motivates a search for features or patterns existing in hospital data collected during triage to identify those children whose fall-related injury results in severe outcome, who need to be retained for further observation in the emergency room, or who must be admitted to a hospital. The current study will develop pattern classification systems to predict the severity of outcome for a patient newly admitted to the emergency room following a fall-related injury and identify features that are associated with severe fall injury to help guide future fall injury prevention programs.

1.2 CHIRPP Database

1.2.1 About CHIRPP

Canadian Hospitals Injury Reporting and Prevention Program (CHIRPP) is a unique national injury surveillance system for the emergency departments of 10 pediatric and 4 general hospitals in Canada [Mackenzie SG et al. 1999]. In the participating hospitals, all families of children presenting to the emergency room due to an injury or poisoning were invited to participate in this program. Once the injured child arrived at a hospital, a research nurse would approach the accompanying adult(s), or the child if possible (i.e. if old enough), to explain the purpose of the

program and to provide them an information sheet. If they agreed to participate after the full explanation and understanding the information sheet, they would be asked to complete a one-page questionnaire including personal information of the injured children and the circumstances of the injury.

Physicians who provided treatment to the injured children are asked to record clinical information of the injury on the back of the same form. The completed form is sent to the Public Health Agency of Canada in Ottawa, where the form is coded and the data is entered into a database. Personal information, such as names and addresses, are removed from the form before it leaves from the hospital. A research unique ID number is added to identify each individual. To guarantee the quality of the data collection, there is a Director who obtains the cooperation of research nurses in the distribution and completion of the questionnaire and supervises the transmission of the forms from the hospital to the Canadian agency. The cumulative national database is kept at the Public Health Agency of Canada.

The primary focus of the program is on childhood injury, although general hospitals also provide services to adults. The wide geographic distribution of participating children's hospitals and the accumulation of CHIRPP database since 1990 make it invaluable for childhood injury research. Therefore, there is an opportunity to find ways to extract the information or patterns embedded in the data to effectively guide future injury prevention programs.

1.2.2 CHIRPP Database Information

The CHIRPP database informs on how, why, and to whom injuries are happening by collecting detailed information on injuries resulting in emergency

rooms visits. Many different variables used to describe the injury information are contained in this database. Some provide the social and demographic information about the injured children such as age, sex, ethnicity and postal code; others describe the circumstances of the injury such as location (the place where the injury occurred), area (the specific part of the place where the injury occurred), context (described by the variable *cntxt*; indicates the activity of the person at the time of injury, such as sports, playing or sleeping), mechanism (described by the variable *mf1;* indicates the type of energy transfer that caused injury, such as a moving swing struck a child running by), and contributing factors (described by the variable *cf1;* indicates the person or object which did not directly malfunction, but made contribution to the injury, such as a table, if a child fell from it). From these variables we are able to identify what activity was taking place before the fall, the location and height of the fall, and the characteristic of the surfaces with which contact was made. In addition, injury date and time, and injury group (described by the variable *injury_gro*; used to identify certain types of injuries such as falls, road traffic accident, abuse, suicide) are also contained in this database. The latter variable (*injury_gro)* was used to extract fall injuries from CHIRPP data. The information of physician's summary about the injury is described by the variables *noi* (nature of injury), *body part* (which part of the body is injured), and *disposition* (treatment of injury, an indicator of severity of the injury).

Values of most of these variables are coded in numbers, although free text accompanied each of the records for reference when needed. Most of the variables in the database are nominal variables, which classify data into categories labeled by

numbers. For example, the variable of *locate* describes where the injury happened and has as many as 53 values represented by unique numbers. Such variables are called multi-nominal variables, whereas variables with only two categories, such as variable *sex,* are called binomial variables (represented by a binary code 1 and 2). There are also a few ordinal variables in this database such as the variable of "*rx*" coded from 1 to 9 indicating the severity of injury from minor to severe. *Age* is the only continuous variable in this database encoding the patient's age in months at the time of the injury. To better understand the CHIRPP database, the main variables used in this study and their corresponding definition are listed in Table 3.

1.3 Thesis Objectives

1.3.1 Development of Logistic Regression, Decision Tree and Artificial Neural Network Models to Predict Severe Fall-related Injuries

Logistic regression (LR) models, decision tree (DT) models, and artificial neural network (ANN) models are developed to predict severe fall injuries. Each model is established using training data and evaluated using test data; therefore, they can be used to classify a new individual patient as to whether he/she is seriously injured or not.

The LR model was used to identify predicting variables that are significantly associated with the severity of childhood fall injury. Only these variables were used to build DT and ANN models to see if their performance could be improved compared with models built using all available variables.

This study used the SAS program (SAS 9.1. SAS Institute Inc.) to build logistic regression models, See5.0 (RuleQuest Research) to build decision tree models, and the feedforward ANN structure (MIRG) trained with the backpropagation algorithm to build ANN models. All models were trained and evaluated using data collected by Canadian Hospitals Injury Reporting and Prevention Program (CHIRPP) [Mackenzie SG et al. 1999].

1.3.2 Comparison of LR, DT and ANN Models

The logistic regression model is a traditional statistical model, whereas DT and ANN are non-linear data mining approaches. This study compares these three models to determine which model performs best for the problem of childhood fall injury severity prediction.

1.3.3 Identification of Important Factors Leading to Severe Fall Injuries

The final objective of this study is to identify the most contributive variables from the best LR model, DT model and ANN model. The identified significant variables will be considered important for childhood fall injury severity prediction and can be used to guide future injury prevention initiatives.

The above-mentioned objectives will be addressed and extensively discussed in the following chapters.

1.4 Contributions to Knowledge

1) This is the first study exploring the severity of childhood fall injuries by using CHIRPP data.

2) LR, DTs and ANNs were first successfully applied and compared in the same study using CHIRPP data. By using dummy variables and selecting the input variables of ANN models by stepwise logistic regression analysis, the ANN models produced promising results. DT models were considered as the most convenient ones among them due to both less work required in data preparation and faster training process.

3) Traditional logistic regression analysis was as good as data mining tools for this research purpose.

4) Develop SAS programs to preprocess the data.

1.5 Thesis Outline

This thesis includes the following six chapters:

Chapter 1: Introduction. This chapter presents the motivation, problem statement, thesis objectives and contributions, and also introduces CHIRPP and the database used in this study.

Chapter 2: Literature review and relevant work. This chapter presents the necessary background information for this thesis.

Chapter 3: Methodology. This chapter details the major calculation steps for training and evaluating LR, DT, and ANN models.

Chapter 4: Results and Discussion. This chapter presents and discusses the results from the above calculations.

Chapter 5: Conclusions and Future work. This chapter summarizes the results and contributions to knowledge made in this study, and provides suggestions for future work.

CHAPTER 2. LITERATURE REVIEW and RELEVANT WORK

This chapter describes the three models - logistic regression models, artificial neural network models and decision tree models. It also reviews some relevant work and applications about the three models in the setting of clinical risk estimation and decision making.

2.1 Logistic Regression

Logistic Regression (LR) can be used to "model the effect of one or several independent variables on the risk of a dichotomous outcome" [D'Agostino RB et al. 2006]. The model can be built using a forward, backward, or stepwise covariate selection process according to the statistical significance test of coefficients for independent variables [Menard SW 2002]. By fitting data to a logistic curve (Figure 1), the predicted probability of the outcome ranges from 0 to 1. LR is widely used in the prediction of medical outcomes. The response variable of LR is usually a categorical random variable with only two alternative outcomes (a binary response) [Pagano M et al. 2000]. Examples of binary response variables include alive or dead, presence or absence, disease or non-disease, success or failure.

2.1.1 Odds and Odds Ratio

In order to understand LR models, we need to introduce odds and odds ratio (OR) first. Odds is a common way in statistics to represent the chances that an event will occur. The odds of an event (outcome) is the ratio of the expected probability that an event will occur to the expected probability that an event will not occur [Allison PD 1999].

$$Odds = \frac{p}{1-p} \quad (2.1)$$

In this study we are trying to predict the occurrence of severe fall injuries by a number of categorical variables. Table 1 below provides a sample calculation for the variable *sex* as an example [Pagano M et al. 2000].

Table 1 Odds and odds ratio

Sex	Severe fall injury	Non-severe fall injury	Total
Boys	a	b	a+b
Girls	c	d	c+d
Total	a+c	b+d	a+b+c+d

The odds for boys to have severe fall injury is $\frac{a/(a+b)}{b/(a+b)}$

The odds for girls to have severe fall injury is $\frac{c/(c+d)}{d/(c+d)}$

The odds ratio of boys to girls is the odds for boys divided by the odds for girls, which is $OR = \frac{a/(a+b)}{b/(a+b)} / \frac{c/(c+d)}{d/(c+d)} = \frac{ad}{bc}$ (2.2)

From equation 2.3, we can get 95% confidence interval of OR.

$$95\%\text{ confidence interval of } \ln(OR) = \ln(OR) \pm 1.96\sqrt{(\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d})} \quad (2.3)$$

If the odds ratio and its 95% confidence interval are greater than 1, we can say that boys are more likely to have severe fall injuries than girls.

2.1.2 Logistic Function

The logistic model is based on a linear relationship between the natural logarithm (ln) of the odds of an event (dependent variable) and one or several independent variables. When there are multiple independent variables involved in the model, it is called multiple logistic regressions. The format of the relationship between dependent and several independent variables is as follows:

$$Ln(odds) = \ln(\frac{p}{1-p}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k \quad (2.4)$$

Algebraic manipulation permits us to express p as:

$$p = 1/(1 + e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+...+\beta_k X_k)}) \quad (2.5)$$

This is called logistic function and its graph is presented as Figure 1. The characteristics of this figure include: 1) the logic function is a common sigmoid curve. The initial stage of growth is approximately exponential; then, as saturation begins, the growth slows, and at maturity, growth stops; 2) X ranges from $-\infty$ to $+\infty$ (negative infinity to positive infinity); 3) Y is in the domain from 0 to 1.

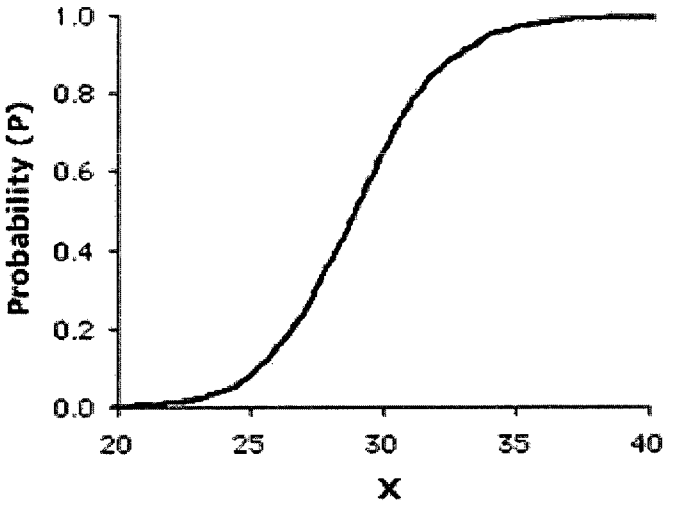


Figure 1 The Logistic function
(Source: http://www.terraseer.com/help/stis/Statistics/Regression /About_Aspatial_Logistic_Regression.htm)

The logistic regression is designed to describe the probability of an event as p, ranging between 0 and 1. In this study, such a probability indicates the risk of a child experiencing a severe fall injury. In Figure 1 illustrating the logistic function, probability(P) is viewed as the probability of an event for a given value of X, while the X is defined as the combined contributions of several risk factors (X_1, X_2, ..., X_k). The characteristics of the logistic function determine the value of p in the function ranging from 0 and 1, which is why the logistic regression is often the first choice when a probability is to be estimated [Dietz et al. 2002, p.6].

In the logistic function (equation 2.5), p is the probability of an event and β_0 is called intercept. The intercept is the value of p when the value of all independent variables is zero. If the model includes all risk factors of event, the intercept will represent the basic probability of an event without any risk factors. β_1, β_2, ..., and β_k are called regression coefficients of X_1, X_2, ...and X_k (where X_1, X_2, ...and X_k are the

independent variables). The regression coefficients show the magnitude of respective contribution of each risk factor to the event probability [Dietz et al. 2002, p.8]. A positive regression coefficient indicates that this independent variable increases the probability of event, while a negative regression coefficient means that this independent variable decreases the probability of event. A larger regression coefficient (absolute value) means that this independent variable has a stronger effect on the probability of the outcome, while a near-zero regression coefficient means that this independent variable has little effect on the probability of the outcome.

In equations 2.4 and 2.5, β_k represents the difference in log odds given a one unit increase in the independent variable X_k. e^{β_k} is the odds ratio for X_k, which is the value that the odds are multiplied by, when the value of the independent variable is increased by one unit. The odds ratio for X_k does not depend on the value of X_k [Larsen. 2008]. The logistic model assumes a nonlinear relationship between the probability of outcome and the independent variables and allows the assessment of interaction between independent variables [Allison PD 1999. chapter 2].

When $\beta_k=0$, $e^{\beta_k}=1$ means that the odds ratio for X_k, equals to 1 and therefore the specific independent variable X_k does not affect the dependent variable. When $\beta_k>0$, $e^{\beta_k}>1$ means that the odds ratio for X_k is larger than 1. If the lower confidence interval of the odds ratio is larger than 1 as well, we conclude that this independent variable X_k increases the odds of an event. When $\beta_k< 0$, $0 < e^{\beta_k} <1$ indicates that the odds ratio for X_k is less than 1.0. If the upper confidence interval of the odds ratio is less than 1.0 as well, we conclude that the independent variable decreases the odds of an event.

The odds ratio and its confidence interval are the most important measures of association from a logistic regression model without any special assumptions, regardless of what the study design is. A fitted logistic model can be used to predict the risk of an individual with specified explanatory variables [Dietz et al. 2002, p.11].

2.1.3 Parameter Estimation for LR Models

Maximum likelihood estimation is a popular statistical method, which has been used for fitting various statistical models to data and providing estimates for the model's parameters [Dietz et al. 2002, p.104]. The values of parameters from β_0 to β_k in a logistic regression model are obtained by maximum likelihood estimation, requiring iterative computational procedures [Statsoft Inc. 2008]. For a fixed data set and determined probability model, maximum likelihood uses the information in a sample to select the model parameter estimates that are most likely to have produced the observed data [Larsen. 2008].

2.1.4 LR Model Creation

The objective of multiple logistic regressions is to discover what combination of explanatory variables provides the best fit for the observed probability. To accomplish this objective, a model is created that includes all independent variables that are assumed to be useful in predicting the response variable. If the number of variables in the model is to be reduced based on significant test, various selection procedures can be employed including forward selection, backward selection and stepwise selection [Cook et al. 2000]. These procedures are used to find the overall best model. In fact, different procedures can arrive at different final models given the

same data. All these procedures are based on purely statistical tests; therefore, the application of prior knowledge into the variable selection process is also of importance. There are typically two goals for regression modeling: one is to obtain a valid estimate of the exposure-outcome relationship. Exposure can be explained by independent variables and outcome is the dependent variable. The other goal is to obtain a predictive model [Dietz et al. 2002, p.165]. Depending on the research objective, different approaches can be applied to obtain the best model.

The forward algorithm is a depth-first greedy search. When the first independent variable is introduced into the model, forward selection examines each independent variable individually and selects the single independent variable that best fits the data. After the first variable is determined, other variables are examined to see how much they can add to the overall fit of the model. Among the remaining variables, the one that adds the most to the overall fit will be introduced into the model. This following step examines remaining variables based on those variables already in the model and introduces those that add significantly to the overall fit; this is repeated until none of the remaining of variables is introduced into the model or there are no variables remaining [Cook et al. 2000].

Backward selection starts with a model containing all of the independent variables available. Each variable is examined to see the effect of removing it from the model on the overall fit. When the removing of a variable leads to the smallest change in the overall fit of the model, this variable is removed. This process continues until all variables in the model have a significant impact on the overall fit of the model [Cook et al. 2000].

A stepwise selection procedure is a combination of the forward and backward selection. It can start either from forward or backward selection. If this procedure starts with a forward selection, after every step, each variable added earlier is examined to see if it is still significant because of the later introduction of variables or combination of variables. Similarly, if the procedure starts with a backward selection, after every step, it will be checked if a variable that has been dropped should be re-added to the model [Cook et al. 2000].

The stepwise regression can be used in the exploratory phase of research, but it is not recommended for theory testing. The main aim of theory testing is to test *a priori* theories or hypotheses of the relations between variables, while the goal of exploratory testing is to discover relationships between dependent and independent variables. In exploratory testing, there is no *a priori* assumption on the relationship between variables [Christersen. 1997, p.232]. Therefore, we choose stepwise regression in this study to create the logistic regression model.

2.1.5 Statistical Significance of Model Effect

The tests for significance of the independent variables in the model can be performed via different techniques.

Wald statistic (test): The Wald statistic is commonly used to test the significance of individual logistic regression coefficients for each independent variable. It aims to test the null hypothesis in logistic regression that a particular β is zero. The Wald statistic is the squared ratio of the non-standardized logistic coefficient to its standard error [Dietz et al. 2002, p.134; Garson GD 2009].

The likelihood ratio test.: The likelihood ratio is a function of log likelihood. Likelihood is the probability that the observed values of the dependent variable may be predicted from the observed values of the independents. While the likelihood varies from 0 to 1, the log likelihood (LL) ratio varies from 0 to minus infinity, because the log of any number less than 1 is negative. LL is calculated through iteration using maximum likelihood estimation (MLE), which is the basis for tests of a logistic model. Because -2LL has approximately a chi-square distribution, the likelihood ratio test is developed based on -2LL (deviance) for assessing the significance of a logistic regression. The likelihood ratio test is a test of the significance of the difference between the likelihood ratio (-2LL) for the researcher's model minus the likelihood ratio for a reduced model. This difference of the likelihood ratio is called "model chi-square." The likelihood ratio test is generally preferred over its alternative - the Wald test [Dietz et al. 2002, p.130 Garson GD 2009].

SAS output gives us results from both of these two methods. The results from these two methods are always consistent.

2.2 Artificial Neural Networks (ANNs)

ANNs use a non knowledge-based type of network to perform nonlinear statistical modeling. There are very few assumptions that need to be verified before neural network model construction; therefore, they are particularly useful for pattern

recognition and data classification when there is a large number of variables, and the relationship between the variables is unknown, nonlinear, or complex – a situation that is difficult to handle using traditional statistical tools [Jaimes F et al. 2005, Tu JV et al. 1996, Hinton GE et al. 1992]

Artificial neural networks work by mimicking human intelligence in machines. They learn from experience, deal with fuzzy situations to abstract essential characteristics from inputs and make generalization to predict unseen data. During the training process, links and layers between neurons are created mimicking the way the human brain processes information. Usually a large number of interconnected neurons work simultaneously to analyze complex relationships between a set of input and output variables. The strength of the connections (weights) between neurons is adjusted by the learning process to reduce the error. The output neuron(s) are a nonlinear combination of the inputs weighted by the parameter weights [Dreyfus G 2005].

A mean squared error classification ANN (MSE ANN) package written in MATLAB (MathWorks Inc) for mining medical data was developed by the Medical Information technology Research Group (MIRG) led by Dr. Frize at Carleton University. This package creates feed-forward ANNs trained via the backpropagation algorithm with automated optimal parameter selection [Frize M. et al. 2000]. MIRG researchers have successfully predicted a variety of clinical and administrative outcomes by using this tool [Frize M et al. 1995]. Some important elements of MSE ANNs used in this study are detailed below.

2.2.1 Network Architecture

A neuron is the individual processing unit in ANNs, which receives input signals and produces output(s). This process is illustrated in Figure 2. First, the neuron combines the input values with weights and bias, and sums them together. The sum is then presented to a transfer function to calculate a numerical activation value, which is passed to other neurons along the connections [Parks RW et al. 1998]. The hyperbolic tangent function is used as transfer function in this study. It is not only flexible and non-linear, but also continuous and differentiable. Compared with other sigmoid transfer functions, it has a faster transfer speed resulting in a faster training process [Bishop 1995]. Furthermore, the output of a hyperbolic tangent function ranges from -1 to 1 with transfer curve across zero, which is especially useful in the case of medical outcome prediction.

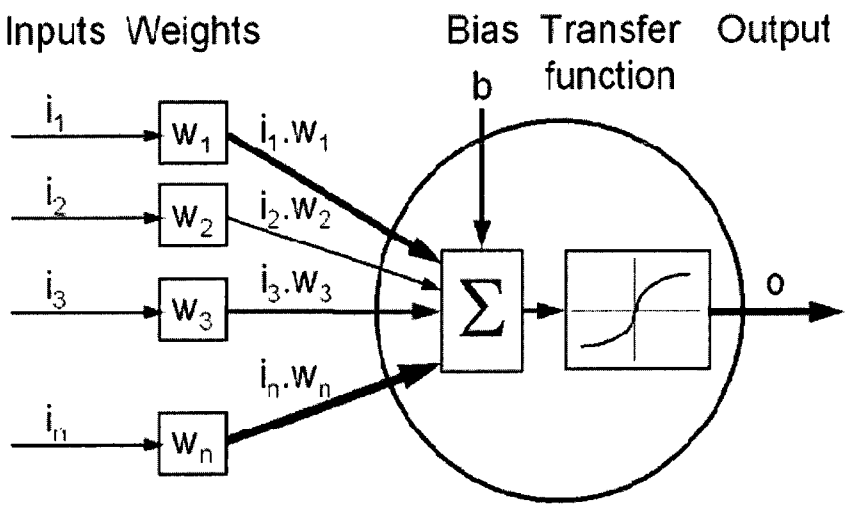


Figure 2 An artificial neuron
[Source: Ray McBride Online 2010
http://raymcbride.com/2010/01/15/artificial-neural-networks]

These processing units are commonly arranged in a series of layers named input layer, one or more hidden layer(s), and an output layer. Neurons in different layers are connected by a set of weights which encode knowledge extracted from the training set and are in charge of transferring signals between neurons in different layers. For a neural net work to be able to learn, a learning strategy (an algorithm) must be specified to guide the learning/training process, from which the link weights are adjusted to improve learning. Supervised and unsupervised learning are mainly two types of learning strategy. Back propagation is widely used supervised learning algorithm, which will be introduced in detail in next section. [Leondes CT et al. 2003].

Besides a learning algorithm, properly designed network architecture is another requirement for a net work to be able to learn. Most neural network applications implement multilayer networks. Feedforward multilayer networks with nonlinear sigmoid transfer function are often termed multilayer perceptrons (MLPs) [Dreyfus G 2005]. In a MLPs, the information flows only from the inputs, hidden layer(s) to outputs. Each hidden unit is connected with every unit at the bottom and upper layers, but units are not connected with others within the same layer. Hidden layers encode the non-linear relationship between the inputs and outputs and form complex decision regions [Leondes CT et al. 2003]. The number of input units at the input layer is determined by the number of independent variables, and the number of units at the output corresponds to the number of classes to be predicted.

Searching for optimal number of hidden units is critical to balance an ANN model's complexity and approximation ability involving many trial and error methods. A higher number of hidden nodes increases model complexity and could result in

overtraining which may lead to poor performance on test data, whereas a lower number of hidden nodes might decrease the model's ability to approximate the patterns of training data. The developed MSE ANN by MIRG applies Kolmogorov's theorem to look for the optimal number of hidden nodes, which states that if there were n input variables, from n to 2n+1 nodes in the hidden layer should be attempted in order to obtain the best possible model [Ennett CM et al. 2004, Rybchynski 2005]. This research applied a 3-layer feed-forward network with one hidden layer and one output shown as below (Figure 3). The activation of the one output node represents the presence of the interested outcome (severe outcome of fall injuries).

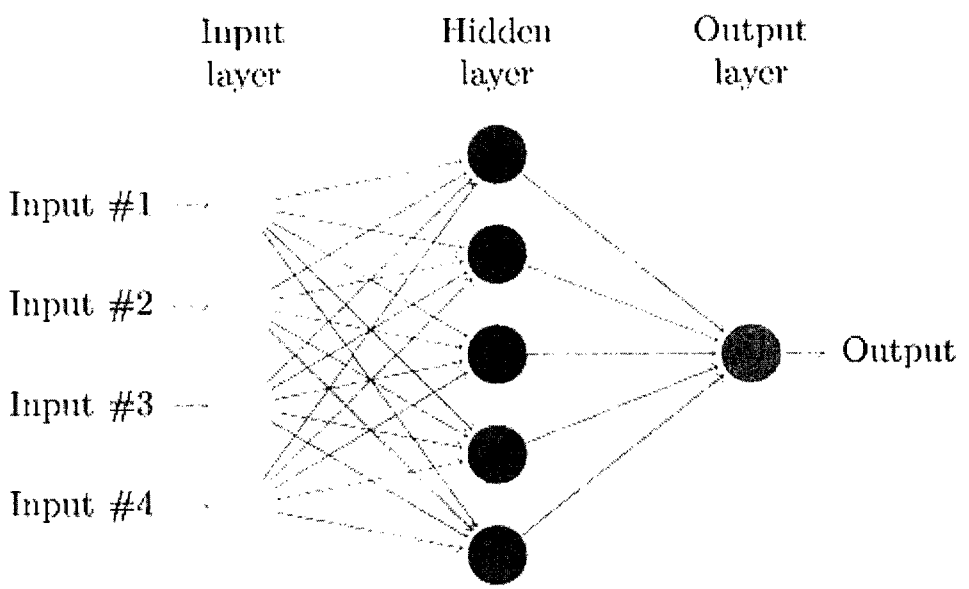


Figure 3 Structure of a 3-layer neural network
[Source: Fauske KM 2006]

2.2.2 Back Propagation Algorithm

During a supervised training process, an ANN applies the input signals in training data to produce output(s), which are compared with target output(s). The

difference between them referred as performance error is calculated. To reduce the error, the training process iterates by adjusting the network parameters such as weights until an acceptable error level is achieved. Therefore, the purpose of training a neural network is to make the actual output(s) approaches the target output(s) by minimizing the error infinitely.

Backpropagation is a very popular learning algorithm for training feed-forward neural networks. At least two layers of weights are required for using this learning algorithm. The basic idea is that the performance error can be greatly reduced by changing the weights. Each iteration of backpropogation algorithm consists of two phases: a signal propagation phase and an error backpropagation phase. Firstly, input layer neurons receive signals which are multiplied by the corresponding weights and summed with bias, then propagated through the hidden layers, to ultimately produce output(s) at the output layer. Second, if the output(s) matches the target output(s), then weights and biases will not be adjusted. If not, error signals will be calculated from the difference between actual output(s) and target output(s) at the output layer and propagated backward to the input layer for adjusting the weights and bias. As a result, the trained network will predict the training data as accurately as possible.

2.2.3 Control Parameters

There are nine control parameters including learning rate (lr), learning rate increment (lr_inc), learning rate decrement (lr_dec), weight-decay constant (λ), weight-decay constant increment (λ_inc) and weight-decay constant decrement (λ_dec), weight-elimination scale factor (wo), momentum (m) and error ratio used to modify the connection weights and biases of the network to improve the model

performance. Among them, the parameters learning rate, learning rate incremental and decrement multiplication factors, momentum and error ratio are always used in the ANN learning algorithm. Weight-decay constant, the weight-decay constant incremental and decrement multiplication factors, and the weight scale are used when the weight elimination cost function is turned on [Rybchynski 2005].

The MSE ANN developed by MIRG implements weight elimination cost function and automatically selects optimal values for the parameters over pre-defined ranges [Ennett CM et al. 2003].

2.2.4 Stopping Criteria - Logarithmic Sensitivity Index

In medical settings, the distribution of the outcome variable is usually highly imbalanced. For example, in CHIRPP database, less than 10% of fall injuries have severe outcome, while the remaining injuries are are minor or moderate. Howver predicting the rare outcome successfully is much more important; therefore, ANN models with higher sensitivity are desired.

Researchers in MIRG developed an automatic ANN tool by using log sensitivity index as a stopping criteria, which can automatically search for a good balance between sensitivity and specificity over the test set while giving more weight to the sensitivity resulting in more successful prediction of the positive outcome (here meaning severe fall injury) [Ennett CM et al. 2002, Rybchynski 2005]. Several stopping criteria were investigated – highest correct classification rate, lowest mean squared error or highest log-sensitivity index value for the automated ANN tool applied in rare outcome medical settings and concluded that the automated ANN tool

could achieve the best classification performance using log-sensitivity index as a stopping criterion [Ennett CM et al. 2004, Rybchynski 2005].

2.2.5 Data Re-sampling and Normalization

Another possible method to deal with the highly skewed distribution of the predicted outcome variable and improve the model performance on the lower prevalence outcome is to re-sample randomly from the low prevalent outcome in the training data until a prevalence of at least 20 percent is reached artificially. The network then has a better chance to learn the patterns attributable to these positive cases resulting in improved performance on test data (testing on the test data is conducted using the actual prevalence). It was suggested by Ennett that the model performance could be improved if the rare outcome could reach 20% [Ennett et al 2000].

MIRG'S ANNs used the hyperbolic tangent function as the transfer function, which requires the input values to be normalized between -1 and 1, and the resulting output ranges between -1 and 1 [Rybchynski 2005].

2.2.6 Feature Selection - Input Variable Reduction

The CHIRPP database was designed to collect detailed information of injuries to serve a variety of research purposes. Some variables contained in this database might be irrelevant to our research purpose meaning some features are noisy and classes in these feature spaces are overlapping. Selecting a set of representative variables which are able to emphasize differences between classes, not only decreases the training time, but also leads to better generalization [Maimon O et al 2005 chapter 5].

Conventional feature selection methods are usually computationally complex, as they need to evaluate many different feature subsets and select the best among them. Many researchers have explored using hybrid systems to improve the performance of data mining techniques, meaning that a particular algorithm was used as a pre-processor to discover significant feature subsets for a primary learning algorithm [Maimon O et al 2005 chapter 5]. Chizi and Maimon introduced that a logistic regression coefficient can be used as a variable selection method for data mining methods [Chizi B et al 2002]. This research applied logistic regression analysis to select important variables, which are used as inputs for neural network model creation.

2.2.7 Model Evaluation Using Verification Data sets

In the ANN model training process, the training set is used to learn by adjusting the weights and bias, and stopping the training depends on the results of the test set [Rybchynski 2005]; therefore, both training set and test set participate in the model creation. A third dataset – the verification set – containing completely unseen dataset with the actual (potentially unbalanced) prior distribution is needed to validate the training results. If the prediction results are good for training and test datasets, but poor for verification sets, the ANN model has most likely over-trained. The model is useful only when its performance on verification sets is acceptable, which indicates that the model generalizes well

In this study, the verification set was created using a secluded section of the original dataset that was never previously presented to the ANN. After the model creation process, the verification set was presented to a single epoch of the ANN

model by "using the weights, bias and network structure corresponding to the best point in the parameter trials" to verify the model performance [Rybchynski 2005].

2.3 Decision Trees

Decision trees (DTs) are popular techniques in supervised data mining due to their simplicity and transparency, which could be used for either classification or regression tasks. Classification trees apply the attribute values to classify an object into a set of predefined classes and are represented graphically as hierarchical structures, from which the relation between the inputs and outputs are easy to be understood and interpreted. DTs can handle both continuous and categorical input attributes naturally [Rokach L et al. 2008].

We used See5 for the decision tree analysis, which is the most recent version of the decision tree algorithm evolved from ID3 and C4.5 developed by an Australian researcher J. Ross Quinlan.

2.3.1 Theory Frame and Evolution History of ID3 Family of Decision Tree Algorithms

The ID3 algorithm is a very simple form, which uses entropy based information gain as the splitting criteria. "Information gain is an impurity-based criterion that uses the entropy measure (origin from information theory) as the impurity measure" [Quinlan 1987]. During the tree induction process, the attribute with the greatest information gain is selected for node splitting. The tree growing stops when all instances belong to a single value of target feature or when best information gain is

not greater than zero [Quinlan 1986]. There are no pruning procedures in ID3. It can not handle numeric and continuous attribute values, and missing values.

C4.5 is another decision tree creation algorithm evolved from ID3, developed by Quinlan in 1993 [Quinlan JR. 1993]. It adds gain ratio as another splitting criteria choice, which is a normalized version of information gain. The impurity-based criterion applied in ID3 is biased towards attributes with larger domain values, which means it prefers input attributes with many values over attributes with fewer values [Quinlan, 1986]. For instance, the attribute of ID number will probably get the highest information gain, because the value of this attribute are always different for each case. However, adding this attribute to a decision tree will result in a poor generalized accuracy. For that reason, the improved criterion of gain ratio is used in C4.5. It has been shown that using gain ratio criteria tends to result in more accurate and less complex classifiers; therefore, it outperforms simple information gain criteria, both from the accuracy and classifier complexity aspects [Quinlan, 1988]. C4.5 can handle both numeric and categorical attributes and even data with missing values. Error-based pruning is performed after the growing phase. See5 incorporates many further improvements, which will be introduced in detail in the following paragraphs.

2.3.2 See5.0 Application

2.3.2.1 Essential Files and Optional Files

See5 applies the values of input variables to predict a case's class by using a user-friendly graphic interface in Windows. A names file and a data file are essential

for the See5 application. The names file clearly describes the types of attributes (variables) and indicates which variables are inputs and outputs. The data type in See5 can be continuous, discrete, ordered discrete, dates, times, case labels. Usually the variable ID is specified as case labels, which is not used to build a classifier. Users can flag attributes as not applicable and define new attributes as functions of other attributes in the names file.

See5 extracts patterns from the training data file. Each column in the data file corresponds to each attribute in the names file sequentially. Missing values are expressed as "?" in the data file. Another file is a test file, which is optional and is used to evaluate the performance of the classifier on unseen test cases. Data in the test file should have exactly the same format as the data file. Another optional file is a cost file, which allows different costs to be defined for each combination of predicted class and real class when some misclassification errors are more severe than others. As discussed below, this cost file provides another way to deal with unbalanced training sets, where missed positives can be assigned a higher cost than false positives to encourage prediction of the underrepresented class.

2.3.2.2 Pruning, Winnow and Boosting

Decision trees are constructed by See5 in two phases: first a large tree is grown to fit the data as closely as possible, and then some sub-trees predicted to have a relatively high error rate are pruned by replacing a sub-tree with a leaf node. An option of "Pruning CF" in See5 is used to decide how much initial tree will be pruned.

The default value 25% of "Pruning CF" option may satisfy most applications. Larger values than 25% would cause less of the initial tree to be pruned and thus produce larger trees, while smaller values result in more pruning. See5 applies error-based pruning to predict error directly from the training data. As a result, the pruned tree will be a small tree with minimal class error.

For some applications with hundreds or even thousands of attributes, See5 can automatically discard marginally relevant attributes, which leads to smaller tree size with potentially higher predictive accuracy. This function is called winnow in See5. Because there are only no more than 20 attributes available in this study, the winnow function seems useless to this study.

Adaptive boosting based on the work of Rob Schapire and Yoav Freund [Freund Y, et al.1999] was incorporated in See5 to improve the performance of a classifier. The idea is that there are a series of classifiers constructed from training data. Every classifier focuses on the cases which were misclassified by the previous classifier. Once all classifiers are trained, the predicted class for each test case is processed by each classifier and voting is used to determine the final predicted class. Quinlan concluded that about 10 iterations is an optimum running number for boosting [Quinlan JR 1996].

2.3.2.3 Feature Selection

In order to design a good classifier, we need to select the right set of features for decision tree analysis. Feature selection methods usually consist of two categories – filter and wrapper [Liu H et al. 1998]. If a particular learning algorithm is used to

identify the best feature subset, it is a wrapper approach. An exhaustive search over the entire feature space is conducted to potentially produce the best classifier, although it is computationally expensive. The filter approach uses an evaluation function to assess the merits of the features directly from the data without use of a learning algorithm. The embedded feature selection method in See5 can be considered as a filter approach [Perner P et al. 2004]. At each level of the tree building process, only one attribute with the highest value of gain ratio is selected to split the data. Feature selection could be considered to be restricted to the root node, because only at this point, the selected feature is important over the entire decision space. After this level, the selected features are thought to be important just for the sub-samples. As a result, the globally optimal feature subset may not be produced in terms of classification accuracy.

Some studies observed that the feature selection strategy of See5 is not optimal [Perner P et al. 2004]; therefore, feature subset selection prior to the learning phase might be a reasonable choice. This study is intended to evaluate the effect of using logistic regression coefficient as feature selection method for decision tree analysis.

2.3.2.4 Performance Evaluation

See5 summarizes the performance of the constructed classifier on training and test data separately in a confusion matrix, which pinpoints the classification results for each class, and evaluates the classifier by tree size, errors and error rate. Tree size is the number of non-empty leaves on the tree. Errors and error rate shows the number and percentage of cases misclassified separately. For applications with many

attributes, See5 lists them in order of importance and shows how much contribution of the individual attributes to the classifier calculated from training data only. For example, if See5 shows the usage of attribute "*sex*" is 90% in our study, it means that the decision tree uses a known value of *sex* when classifying 90% of the training cases to create the tree. Higher percentage indicates more contribution in class prediction and thus this attribute may be more important than other attributes with lower percentages.

Confusion Matrix The confusion matrix provided by See5 and shown in Table 2 provides complete information about the performance of the classifier. Predictive accuracy over the test set is usually used to measure the performance of a classification model. It shows how well the designed model will work on future unseen data [Maimon O, et al. 2005]. Models with high accuracy will result in a lower error rate for test data or future unseen data. In confusion matrix shown by Table 2, true indicates this case was correctly classified, like TP (True Positive) or TN (True Negative). TP means positive cases which were correctly classified as positive, and TN means negative cases which were correctly classified as negative. False indicates this case was incorrectly classified, like FP (False Positive) or FN (False Negative). FP means negative cases were incorrectly classified as positive, and FN means positive cases were incorrectly classified as negative. Predictive accuracy also named correct classification rate (CCR) refers to how many cases both negative and positive were correctly classified by the classifier, and is defined as (TP + TN)/(TP + FP + TN + FN), which is consistent with error rate See5 provided.

Table 2 Confusion matrix

	Predicted class			
		Yes	No	
True class	Yes	True positive (TP)	False negative (FN)	TP+FN
	No	False positive (FP)	True negative (TN)	FP+TN
		TP+FP	FN+TN	Total cases

However, sometimes a classifier with higher accuracy does not mean it is more useful, especially for imbalanced data. In our application, we are trying to predict positive (severe) and negative (non-severe) outcomes of falls from a set of input attributes. From the results of our preliminary analysis, there were about 93% negative cases and 7% positive cases and therefore the database was imbalanced. A simple classifier with apparently high accuracy would likely classify all the cases as the majority class, so the negative cases would be classified correctly. However, this classifier is useless, as it can not separate positive cases from negative cases. In this case, sensitivity and specificity would be better to be used as binary classifier evaluation criteria [Han J, et al. 2001].

Sensitivity assesses how well a classifier recognizes positive cases from the total and is defined as TP/(TP+FN), which is also called true positive rate. A constructed decision tree model with higher sensitivity will correctly identify more severe cases. Specificity assesses how well a classifier recognizes negative cases from the total and is defined as TN/(FP+TN), which is also referred as true negative rate. A constructed decision tree model with higher specificity will correctly classify more non-severe

cases. We usually need to use both sensitivity and specificity to evaluate how well a classifier is.

Error Rate and Misclassification Cost

As mentioned above, error rate is the percentage of cases misclassified, which is a measure of accuracy of a classifier. Traditional learning algorithms always aim to reduce the error rate and thereby improve accuracy. However for imbalanced data such as our application, the negative cases occur much more often than positive cases. Raising the cost of misclassifying the minority class is one way to correct imbalance and improve the model.

Variable misclassification cost is a new function incorporated in See5 compared with C4.5. This option is used when some classification errors are more severe than others in some practical applications. Although we rarely know what the individual misclassification costs actually are, See5 can treat different classification errors unequally by defining a separate cost for each predicted and actual class pair. As a result, the constructed classifier will make the expected misclassification cost minimum rather than the error rate. Misclassification costs can be manipulated to achieve the best model.

From the confusion matrix, we can calculate misclassification cost directly. Given the cost of a false positive as C_{FP} and a false negative as C_{FN}, the total misclassification cost is defined as ($FP^* C_{FP} + FN^* C_{FN}$) [Bradley AP et al. 1997]. See5 automatically provides the average misclassification cost, which is the total misclassification cost divided by the number of cases.

The Receiver Operating Characteristic (ROC) Curves

ROC curves are commonly used in medical decision-making [Swets JA et al. 2000], and recently have been increasingly adopted in machine learning and data mining, as researchers have realized that the simple metric of classification accuracy alone is not enough for model performance measurement. ROC curves summarize the performance of a classification model and can be used to evaluate and compare the performance of different classification models. They are especially useful when the data distribution is imbalanced and problem-dependent misclassification costs have to be taken into account [Fawcett T. 2006].

The Receiver Operating Characteristic (ROC) curve is a plot of a model's sensitivity versus its false positive rate (i.e. one minus its specificity). In other words, in the two-dimensional ROC graphs, X axis represents false positive rate, which is FP/(TN+FP), and Y axis represents true positive rate, which is TP/(TP+FN). Moving along the ROC curves from the bottom-left corner to the top-right corner represents a range of trade-offs between true positive and false positive rates [Swets JA. 1988]. Therefore, researchers can visualize the performance of the model over entire range, and thus select a suitable operating point with reasonable sensitivity and specificity according to ROC curves.

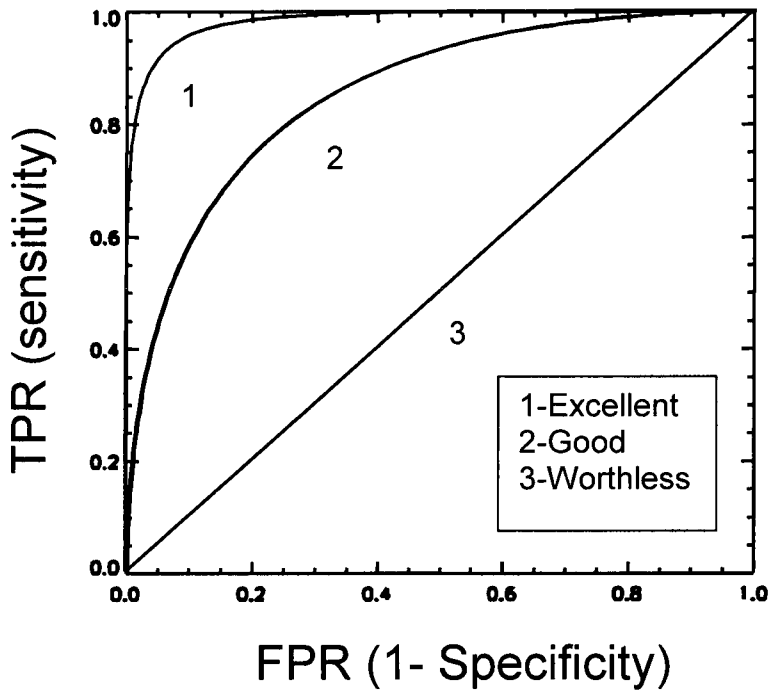


Figure 4 Comparing ROC curves

As shown in Figure 4 about ROC curves, the diagonal represents a random classification model just like complete guessing. The more the ROC curve bulges toward the top-left corner, the better the model separates the target class from the background class. The point (0, 1) represents the best possible prediction model with 100% sensitivity and specificity.

Area under the ROC is a useful metric for classifier performance as it is independent of the decision criterion selected and prior probabilities. It can take on

any value between 0 and 1. The closer AUC is to 1, the better the model is. The random guessing produces a diagonal line, which has AUC of 0.5. Any realistic model should have AUC larger than 0.5. As AUC aggregates performance of a model across the entire range of trade-offs, the AUC comparison establishes a dominance relationship between models.

Another attractive property of ROC curves is that they are insensitive to class distribution changes. That is the ROC curves will not change even if the true positive rate (priori probability) in the test data is different from training data, whereas accuracy will change along with the class distribution changes.

A classifier using decision tree analysis is designed to produce a class label for each case instead of probabilities. Such a discrete classifier applied to test data could only yield a single confusion matrix, which in turn corresponds to a point in ROC space. Varying the misclassification costs, we can get a series of points which constitute a ROC.

LR and ANN naturally yield a probability that represents the degree to which a case is a member of a class. The probability, which is a continuous numeric value, can be used as a threshold to produce a discrete classifier. Each threshold value produces different confusion matrices corresponding to different points in ROC space. A curve made up with these points is a ROC curve. This is the basic idea of ROC curves.

We applied the commonly accepted measures of accuracy along with sensitivity and specificity on the test data set, as well as the area under ROC curve as major model evaluation criteria in this research.

2.4 Application of LR, ANNs and DTs in Medical Outcome Prediction

Although the development of an ANN model is time consuming and computationally intensive, the potential for higher classification accuracy and discriminating power makes it popular in medical data applications. ANN models have been successfully applied in the prediction of clinical outcomes, such as pediatric trauma mortality prediction [DiRusso SM et al. 2002], pediatric meningococcal disease outcome prediction [Nguyen T, et al. 2002], surgical decision making on traumatic brain injury patients [Li YC, et al. 2000], and breast cancer survivability prediction [Delen D et al. 2005]. They have also shown to be efficient in diagnosing thyroid disease [Zhang G et al. 1998] and Methicillin-Resistant Staphylococcus Aureus [Shang JS et al. 2000].

DiRusso et al [DiRusso SM et al. 2002] have compared ANNs and LR in childhood injury analysis. That study developed an outcome (death) prediction model for injured children based on the large National Pediatric Trauma Registry database, which includes all trauma patients who were less than 20 years and admitted to hospital with a primary diagnosis of injury, and covers the majority of child trauma treatment centres in the United States; therefore, it can represent the general pediatric trauma population. Anatomic and physiologic measures were used as predictor variables such as age, sex, systolic blood pressure, heart rate, the New Injury Severity Score, and the Revised Trauma Score. They concluded that their ANN model outperformed the LR model in both discrimination and calibration. The robust performance of ANN was attributed to the ability to detect complex,

multidimensional, and nonlinear relationships between variables and also due to generalization from the noisy or incomplete data.

Eftekhar et al applied an ANN model to predict the mortality in head injury patients [Eftekhar B, et al. 2005]. That study was based on clinical data collected from emergency departments of six major university hospitals in Tehran from 1999 to 2000 including demographic, pre-hospital and in-hospital information. The developed ANN model was significantly better than LR in AUC, whereas LR achieved a higher accuracy. Therefore, the authors suggested that techniques incorporating features of both LR and ANNs might be used to develop the optimal models for injury outcome prediction [Eftekhar B et al. 2005].

LR and DTs have been used for medical outcome prediction as well [Perlich C et al. 2003]. For example, Samanta et al combined the advantage of LR with DTs to generate decision rules for periventricular leukomalacia (PVL) prediction in neonates with complex heart disease using blood gas analysis data [Samanta B et al. 2009] . They applied LR to select the significant variables for PVL occurrence, which were used as DT inputs to build classification rules. The performance of decision rules with and without LR preselected features was compared. They concluded that the decision rules with reduced features preselected by LR had better classification success and greater area under ROC curve. The increased performance was partially due to the fact that the features selected by LR are not correlated to each other as tested by correlation analysis.

Many researchers suggested that data mining and statistical methods were complementary. They attempted using both of them in different settings of medical

outcome prediction. For example, Samanta et al combined the advantage of LR with DTs to generate decision rules for periventricular leukomalacia (PVL) prediction in neonates with complex heart disease using blood gas analysis data [Samanta B et al. 2009] . They applied LR to select the significant variables for PVL occurrence, which were used as DT inputs to build classification rules. The performance of decision rules with and without LR preselected features was compared. They concluded that the decision rules with reduced features preselected by LR had better classification success and greater area under ROC curve. The increased performance was partially due to the fact that the features selected by LR are not correlated each other as tested by correlation analysis.

Decision trees and LR have also been applied to feature-selection prior to training an ANN. Dybowski et al built an ANN model to predict outcome (death) for patients in ICU [Dybowski R et al. 1996]. LR and DTs were used for feature selection, in which significant variables more likely to influence outcome were selected to build the ANN model. They concluded that the developed ANN model was more accurate than the LR model in hospital mortality prediction

In another study [Ge G et al 2008], decision trees were used to classify mass-spectrometry data generated from premalignant pancreatic cancer. The Student's t-test and Wilcoxon rank sum test were used for feature selection, which were shown to be robust in searching for meaningful biomarkers in pancreatic cancer development and for classification purpose. Samanta et al got easily interpretable rules for periventricular leukomalacia prediction by combining decision trees and logistic regression analysis using the postoperative hemodynamic and arterial blood gas data.

Logistic regression models are first used to identify the significant variables, which are used as input features for decision tree models.

MIRG researchers have successfully applied a three-layer feedforward backpropagation ANN tool to estimate a variety of medical outcomes in complex medical settings, such as estimation of in-hospital mortality, length of stay and duration of artificial ventilation for both adult intensive care unit (ICU) patients [Frize et al. 1995] and neonatal intensive care unit (NICU) patients [Walker et al. 1999, Tong Y et al. 2002]. The developed ANN model measured by correct classification rate and average squared error worked well for both ICU and NICU patients. This model classifies patients into two classes – whether the required length of ventilation was eight hours or less vs. more than eight hours. They also proved that the weight-elimination cost function could be used to improve the ANN performance in mortality prediction for coronary artery bypass grafting surgery patients [Trigg, 1997, Frize et al. 1995].

Shi applied feedforward back propagation ANNs to predict repeat childhood injuries based on the CHEO-CHIRPP database from 1990 to 2001 [Shi YP 2004]. The optimal model could correctly predict 80 percent of patients who may suffer repeat injuries. Shi also identified the relative importance of the variables in the ANN model, which might provide useful information for health professionals about how to effectively prevent repeat injuries. Shi suggested that LR models could be developed for the same problem and be compared with the ANN models. In her thesis, the nominal variables were treated as continuous variables and were normalized by subtracting the mean and dividing by a number of standard deviations to reduce the

input values between -1 and 1 (Assuming a Gaussian distribution, this only gurantees that 68% of the data will fall within -1 and 1). As discussed below, the categorical variables may also be encoded using dummy variables.

Erdebil developed an ANN model to predict the severity of all-terrain vehicle (ATV) injuries using a subset of the CHIRPP database [Erdebil Y et al. 2005]. The optimal performance of the model had 47.3% sensitivity, 80.8% specificity, and an area under the ROC curve of 0.711. Erdebil also identified the contribution of individual predictor variables in the model. However, the sample size in her study was relatively small, including only 2,927 patients in total. This small sample had to be divided into training, test and validation data. Previous studies have found that as the sample size of training data is decreased, the performance of the model on the validation set tends to deteriorate rapidly, which means that the performance of ANN models depends partly on appropriate sample size [Clermont G et al. 2001].

In CHIRPP, most independent variables are nominal with no natural order. A series of dummy variables based on the original values can be created and introduced into the ANN model. Each dummy variable takes the value "-1"or "1" according to the corresponding nominal value which is true or false. As a result, the number of inputs would increase by replacing one nominal variable with several dummy variables (one for each possible value). This increases model complexity and may lead to reduced performance over independent test data [Hand DJ 1981]. In this study, feature selection was used to reduce the number of input variables and thus model complexity.

Data mining techniques and statistical approaches can be complementary, rather than competing methodologies for medical data analysis. For example, ANNs can be considered as nonlinear generalization forms of LR [Dreiseitl S et al. 2002]. There are no universal rules for selecting which methods to employ in medical data analysis, and method selection should depend on prior information about each specific problem. Furthermore, it has been suggested that a complementary method combining the features of statistical approaches and data mining techniques would improve the performance on medical data prediction and classification [Paliwal M et al. 2009, Samanta B et al. 2009, Eftekhar B et al. 2005]; therefore, it is worthwhile for us to apply both approaches to analyze childhood fall injuries using CHIRPP database, which is an important and unique source of childhood injury information in Canada.

CHAPTER 3. METHODOLOGY

This chapter describes the methodology used in this research. It covers the following topics:

1. The study dataset was split into training and test sets for LR model and DT model creation, and training, test and verification sets for ANN model creation. The original dataset was handled using the Statistical Product and Service Solutions (SPSS) software package.

2. LR, DTs and ANNs were applied respectively to create models to predict the severity of childhood fall injuries based on records for children newborns up to 14 years of age from the CHIRPP data.

3. The results of the three models were compared in terms of sensitivity, specificity and the area under ROC curve.

3.1 Database Handling

3.1.1 Database Pre-Processing

The dataset of childhood fall injuries used in this study was a part of CHIRPP data in 2007. There were 40,399 cases of fall injuries for children less than 180 months old across Canada in 2007. In this study, we restricted analysis to unintentional fall injuries, so 40,285 observations were chosen as our study subjects.

The initial dataset was in Microsoft Excel format. SPSS was used to remove some variables with missing more than 40% values. The variables that we selected for further analysis and their distribution of values are listed in Table 3.

Table 3 List of variables used in this study
(CHIRPP, 0-14years of age, falls, 2007)

Variable	Description	Variable Type	No. of Unique Values	No. and % of Records With Missing Values
Injury_event_id	Injury event ID.	Nominal	40,399	None
Sex	Patient's gender	Binary	2	None
Age (month)	Patient's age in month	Continuous	0-179	None
Mon	Month of injury date	Ordinal	12	None
Day	Day of injury date	Ordinal	7	None
Hour	Time of injury	Ordinal	24	5,248 (13.0%)
Locate	Injury location (place)	Nominal	53	1 (0.0%)
Area	Specific injury location	Nominal	43	2 (0.0%)
Cntxt	Activity prior to injury	Nominal	45	3391 (8.4%)
Mf1	Mechanism of injury	Nominal	350	6 (0.01%)
Cf1	Contributing factor	Nominal	361	12,085 (29.9%)
Follow up_flag	Agree to follow up?	Binary	2	None
Ni1	Nature of injury	Nominal	31	1(0.0%)
Bp1	Injured body part	Nominal	31	1(0.0%)
Disposition	Treatment received (indicator of the severity of the injury)	Ordinal	11	2(0.0%)

3.1.2 Outcome Variable Definition

In CHIRPP, the ordinal variable of *disposition* indicates the severity of the fall injury, which was used as the outcome in this study. The distribution of *disposition* is shown in Table 4. There were six different levels of *disposition* for injured children in emergency departments according to the severity of fall injuries. Based on the suggestions from CHIRPP experts, the values of *disposition* from 5 to 10 were normally considered as severe injuries, the values of 3 and 4 as moderate injury, and the values of 1 and 2 as minor injury. In this study, severe fall injuries accounted for 7% in all fall injuries, moderate fall injuries for 23%, and minor injuries for 70%, as shown in Table 4.

For the children assigned to the severe injury group, they needed to be admitted to the hospital or remain under observation for a longer time in the emergency department, whereas the moderate and minor fall injured children could be discharged immediately after on-spot treatment in the emergency department or after medical advice. As our research purpose was to predict severe fall injuries, the moderate and minor fall injuries were collapsed into a single group (from value 1 to 4 in Table 4), which were considered non-severe fall injuries. The other two groups (value 5 and 6-10 in Table 4) were collapsed together and considered as severe fall injuries. Therefore, there were two levels in the study outcome, with 1 and 0 indicating severe and non-severe fall injuries, respectively as shown in Figure 5. About 7.35% fall injury cases were classified as severe fall injuries.

Table 4 Definition and frequency distribution of outcome variable 'disposition'

Values	Definition	Frequency	%
1	Left without being seen	827	2.05
2	Advice only	8,378	20.80
3	Treated, follow-up if necessary	13,050	32.39
4	Treated, follow-up required	15,064	37.39
5	Prolonged observation in Emergency	930	2.31
6-10	Admission to hospital	2,036	5.04
Total		40,285	100

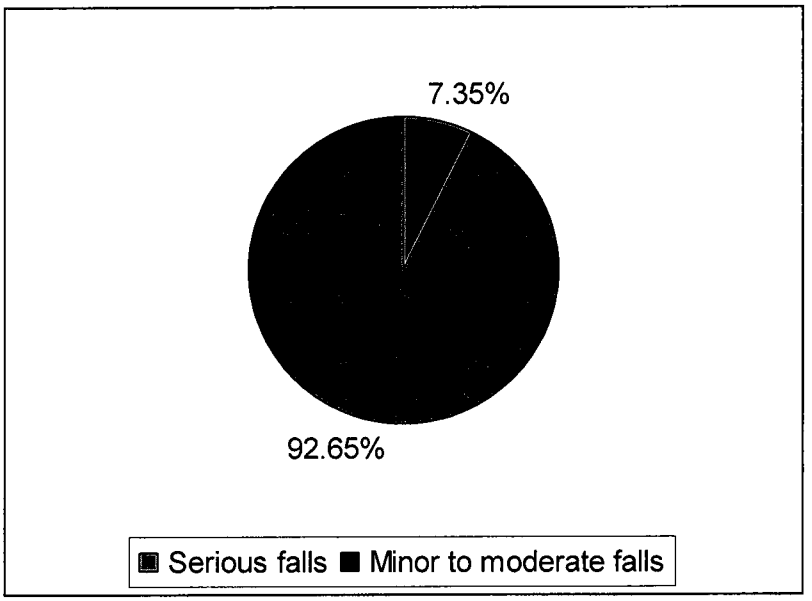


Figure 5 Composition of fall injury severity

3.1.3 Generation of Training and Test Data Sets

The database with selected variables for analysis was split into training set and test set to establish the predictive models and evaluate the performance of models using a hold-out test strategy. The training set included 70% of the fall injury cases, and the remaining 30% of cases were used as the test set. We applied SPSS software to split the data by randomization. In total, there were 40,285 cases for the final analysis. Finally, 28,316 cases were assigned into the training dataset and 11,969 cases into test dataset. The positive rate (priori probability) of the outcome variable (*disposition*) was similar between the training dataset (7.37%) and the test dataset (7.34%), as shown in Table 5. LR and DT model creation used the same training set and test set created in this section. As additional verification data sets were required for neural network model creation, the raw database was split in a different way for ANN analysis described in the ANN section.

Table 5 Distribution of cases in training set and test set

	Training set	Test set	Total
Severe cases	2,088 (7.37%)	878 (7.34%)	2,966 (7.36%)
Non-severe cases	26,228 (92.63%)	11,091 (92.66%)	37,319 (92.64%)
Total cases	28,316	11,969	40,285

3.2 LR Model Creation

A LR model was created using SAS program (SAS 9.1. SAS Institute Inc.). Training data was used to create the model, and test data was used to validate the

established model. The outcome (*disposition*) was defined as 1 when the fall injury was severe, and 0 otherwise. The LR model aimed to predict those fall injuries that resulted in severe outcome (i.e. variable *disposition* equals to 1) using a number of independent variables included in Table 3.

3.2.1 Value Combination and Dummy Variable Creation

The original values of the multinomial and ordinal variables with more than 2 levels could not be used directly for the LR model creation. Most of the nominal and ordinal variables in the original dataset listed in Table 3 had too many values and the sub-sample size for one specific value might be too small. Therefore they needed to be combined into fewer values to fit LR model. Values were collapsed based on similar prevalence of severe outcome. For example, there were as many as 53 distinct values for the original variable *locate,* indicating injury location, including "Own home", "Other home", and "Residential institution", etc. As the rate of severe outcomes was similar between "Own home" and "Other home", they were collapsed as one category. All of the other values were collapsed in this way. As a result, 5 categories were used to represent the original 53 categories for the variable of *locate*. Generally, 4-6 categories were created for each nominal or ordinal variable shown in Table 6.

Encoding categorical variables using numeric values may assign unintentional meanings to the variable values since there was no real numerical relationship between different values. For example, the values of 19, 29 and 39 in *locat*e were assigned as "Own home", "Other home" and "Residential institution" respectively in

the original data. It did not make sense to interpret the values as meaning that "Residential institution" was two times as "Own home".

The solution to this problem was to use a number of dummy variables [Gordon RA 2010]. A value of 1 meant something was true; 0 meant something was false. For example, for the combined variable of *locate*, it was reasonable to create a dummy variable called "Fall in house and apartment". If the fall happened at house and apartment, it was assigned as 1 in this dummy variable; otherwise as 0. This rule can be applied to other categories of combined *locate* variable.

Usually n-1 dummy variables were necessary to represent the original nominal variable with n values. Each dummy variable took two values of 0 and 1. If all dummy variables took the value of 0, the category that was not coded as true. This last category was used as reference group in the LR model. Generally, the reference group was the category with the lowest prevalence of severe outcome. However if the group with the lowest prevalence of severe outcome was "others", we would use the second lowest category as reference. That was because we would not get a meaningful conclusion if we compared a specific category with "others".

Table 6 describes the grouping results and the prevalence of severe outcome of fall injuries in different categories for each variable for the training data set. The category with the lowest rate of severe outcome of fall injury (rare outcome) was usually used as a reference group for stepwise LR analysis. Other categories were compared with the reference group. Dummy variables were created for multi-nominal and ordinal variables with more than 2 levels. All dummy variables were indicated in

this table as Xn-m, where n indicates the order of prediction variables and m indicates the mth category of this variable.

Take the variable locate having 5 values as an example in Table 6. Four dummy variables were created for this collapsed variable. If a fall injury happed at "House", the dummy variable $X_{7\text{-}1}$ would take the value of 1, and the other three dummy variables $X_{7\text{-}2}$, $X_{7\text{-}3}$ and $X_{7\text{-}4}$ would take the value of 0. In this case, "School" was used as a reference group. If a fall injury happened at "School", all of the four dummy variables would take the value of 0. That was why 4 dummy variables could be used to represent one original variable with 5 categories.

If the dummy variable $X_{7\text{-}1}$ was found to be significant in the LR model (i.e. both the OR value and its 95% confidence interval of this variable were larger than 1), this would mean that fall injuries that happened at "House and apartment" were significantly more likely to be severe than those that happened at "School" (reference group). The significant variables could be identified by the values of OR and its 95% confidence interval from LR models.

We developed SAS programs to conduct value combination and dummy variable creation. There were 46 dummy variables created in total.

Table 6 Variables used in LR analysis

Variable		Category	Sample size	Prevalence of rare outcome (%)
Sex	X_{1-1}	Boys	16,574	7.88
		Girls*	11,742	6.66
Age	X_{2-1}	<1	2,035	9.29
		1*	3,107	4.76
	X_{2-2}	2-4	6,545	5.81
	X_{2-3}	5-9	7,795	8.67
	X_{2-4}	10-14	8,834	7.87
Mon	X_{3-1}	1-3	6,112	6.81
	X_{3-2}	4-6	8,498	7.66
	X_{3-3}	7-9	8,099	8.26
		10-12*	5,607	6.28
Day	X_{4-1}	Monday	4,188	7.62
	X_{4-2}	Tuesday	3,997	7.06
	X_{4-3}	Wednesday	3,916	7.48
		Thursday*	3,936	6.73
	X_{4-4}	Friday	4,008	7.24
	X_{4-5}	Saturday	4,069	7.13
	X_{4-6}	Sunday	4,202	8.31

Table 6 Variables used in LR analysis (continued)

Variable		Category	Sample size	Prevalence of rare outcome (%)
Hour		1am-6am*	378	4.76
	X_{5-1}	7am-12pm	6,059	6.44
	X_{5-2}	1pm-6pm	11,401	7.32
	X_{5-3}	7pm-12am	10,478	8.07
Followup_flag		Permitted*	13,076	5.52
	X_{6-1}	Refused	15,240	8.96
Locate		1.School*	4,549	5.76
	X_{7-1}	2.House and apartment (own home and other home)	12,926	6.41
	X_{7-2}	3.Public park and road; other facility for land and ice based sport; camping ground	4,488	9.63
	X_{7-3}	4.Unknown location	3,367	11.05
	X_{7-4}	5.Others	2,986	6.46
Area		1.Bedroom, living room, family room and recreation room*	4,281	4.86
	X_{8-1}	2.Sports related areas	2,572	7.15
	X_{8-2}	3.Playground, garden and yard	5,038	8.24
	X_{8-3}	4.Unknown area	9,860	8.71
	X_{8-4}	5.Others	6,565	6.43
Context		1.Walking, running, crawling*	4,854	5.07
	X_{9-1}	2.Playing, climbing , dancing	9,474	6.99
	X_{9-2}	3.Sports and physical recreation activities	6,288	7.90
	X_{9-3}	4. Others	7,700	8.87
Mf1		1.Stairs or steps (structural element)*	1,795	6.18
	X_{10-1}	2.Concrete and other human-made surfaces; Floors or flooring materials (structural element)	9,137	7.55
	X_{10-2}	3.Ice, snow, frost (environmental element)	2,087	8.86

Table 6 Variables used in LR analysis (continued)

Variable		Category	Sample size	Prevalence of rare outcome (%)
Mf1	X_{10-3}	4. Ground and other natural surfaces; Indoor or outdoor unknown surface (natural environment)	6,669	9.94
	X_{10-4}	5. Others	8,628	5.09
Cf1	X_{11-1}	1.Unknown factors	8,452	4.44
		2.Beds, sofas, couches, divans, chesterfields (furniture) ; Child other than victim (person)*	5,217	6.44
	X_{11-2}	3.Bicycles and bicycle parts and accessories; Monkey bars, jungle gyms and other climbers (playground equipment)	3,408	11.27
	X_{11-3}	4.Others	11,239	8.84
Ni1		1.Open wound*	5,409	1.15
	X_{12-1}	2.Superficial; sprain or strain; soft tissue injury	7,086	1.55
	X_{12-2}	3.Minor head injury	4,167	7.42
	X_{12-3}	4.Fracture	8,596	13.41
	X_{12-4}	5.Dislocation; injury to nerve, blood vessel or internal organ; traumatic amputation; burn or corrosion; poisoning; drowning; concussion; intracranial injury; foreign body in alimentary tract; multiple injuries	1,031	41.32
	X_{12-5}	6. Others	2,027	1.38
Bp1		1.Face; foot*	4,989	0.86
	X_{13-1}	2.Wrist; hand; finger or thumb; clavicle; shoulder; ankle; knee;	6,306	1.97
	X_{13-2}	3.Specified head injury	4,019	7.54
	X_{13-3}	4.Forearm; lower leg	6,812	11.05
	X_{13-4}	5.Upper arm; elbow	2,995	16.56
	X_{13-5}	6.Others	3,195	11.55

* Reference category for LR model creation

3.2.2 Stepwise Logistic Regression Model Creation

The stepwise multivariate logistic regression analysis was used to create the prediction model. The initial regression model included all variables listed in Table 6. All of the variables were entered into the model as dummy variables.

As SAS output provided OR, 95% confidence interval of OR and p value for each dummy variable, from which significant variables could be identified. If the 95% confidence interval of OR was less than or more than 1.0 but excluded 1.0, which was consistent with p value less than 0.05, this variable would be identified as significant. If there was at least one dummy variable originating from the same original variable that was significant, we considered that original variable as a significant variable. If none of the dummy variables were significant, we considered that original variable as non-significant. The non-significant variables were removed from the final LR model. All of the LR models here were created with training data only.

The β values we obtained from the training data were applied to the test data for model performance evaluation. SAS automatically provides a "c value" to indicate the area under ROC curve (AUC). SPSS could be used to plot the ROC curves based on the data from SAS output and to calculate the 95% confidence intervals of the AUC. The AUC and its 95% confidence intervals were compared with other models to evaluate the performance of models.

3.3 ANN Model Creation

3.3.1 Data Normalization

The ANN tools developed by MIRG were used in this study to create double-layered feedforward ANNs trained with the backpropagation algorithm, while these tools had been effectively used to predict various clinical outcomes in other studies. It was interesting to see whether the ANNs were suitable for CHIRPP data analysis, which was a large-scale data set composed of various categorical variables and to see if the dichotomous output – whether the fall injury would be severe or not, was a good fit for the MIRG ANN tools.

As required by the ANN tools, all input data needed to be normalized between -1 and 1. For continuous variables, each value was shifted by the mean and then divided by three times the standard deviation. This method removes outliers and increases the uniform distribution of the data by using statistical measures of mean and variance [Williams S. et al]. There was actually only one continuous variable *age* in the data set, which was suitable for this normalization method. All other variables were categorical and required a different preprocessing approach as described below.

In the preliminary analysis, we tried to normalize all the multi-nominal variables in this way. The means and standard deviations were calculated from the original values of the nominal variables. All the variables shown in Table 3 were normalized in this way and used as inputs to the ANN tools. The preliminary result showed that the performance of ANN models using nominal variables in this way was worse than the LR and DT models.

In order to improve the performance of the ANN tools, we tried to use dummy variables to take the place of the original nominal variables as inputs. As a result, the ANN performance was significantly improved. The best performance of ANNs could

be reached by using the significant variables selected by the final LR model. In total, there were 8 variables used as inputs for ANNs (Table 13), which were significantly associated with the severe outcome of fall injury according to the results of the final LR model; the variable *age* was treated as a continuous variable and was normalized by the method mentioned above. Variable *sex* and *follow-up flag* were binary variables. We used -1 to represent one feature, and 1 to represent the other feature. For the other 5 nominal variables, we used dummy variables which were the same as those used in LR model creation to replace the original nominal variable. For each dummy variable, we used -1 to represent the absence of a feature and 1 to represent the presence of a feature. Finally, we had 24 variables as ANN inputs shown in Table 13.

3.3.2 Data Splitting

After all of the variables were scaled from -1 to 1, the original data was split into three sets: training, test and verification dataset [Rybchynski 2005]. First, 1/3 of the total cases were randomly assigned into a verification set. Second, the remaining 2/3 of the total cases were split into two datasets, with 2/3 of them into training set and 1/3 into test set. Finally, ten subsets of the verification set were created using random sampling allowing for some overlap between subsets. These verification sets contain totally unseen data so as to be used to validate the performance of the trained ANN model.

3.3.3 Data Re-sampling

As mentioned previously, the severe outcome of fall injuries only accounted for about 7% of the total fall injuries. In order to increase the severe outcome up to 20%, we made three copies of each positive case in the training data only. The resulting number of cases in each data set is shown in Table 7 below; 22,894 training were used as inputs to the ANNs. The ANN tool automatically tried structures with zero to 2n+1 hidden nodes to optimize the performance, where n is the number of input variables.

Table 7 Number of cases in training, test and validation sets for ANNs analysis

	Training	Test	Validation
Number of Cases	22,894	8,376	11,969
Prevalence of severe outcome (%)	23.9	7.43	7.44

3.3.4 Defining the Default Value and Range for Each Parameter

There were nine parameters that affect performance in the ANN model training process. Their default values and ranges shown in Table 8 were determined previously by MIRG researchers, who had tried this tool on a number of different medical outcome predictions using a number of medical databases including the Canadian Neonatal Network Database, the Evidence-based Practice Identification and

Change Database, and the Children's Hospital of East Ontario Database [Townsend D 2007, Qi L 2005, Rybchynski 2005, Ennett CM 2003].

Table 8 Default parameter range: minimum, maximum and start points

Parameter	Start Point	Minimum	Maximum
Learning Rate	0.0005	0.00005	0.01
Learning Rate Increment	1.001	0.75	1.25
Learning Rate Decrement	0.999	0.75	1.25
Lambda	0.0001	0.00001	0.0001
Lambda Increment	1.001	0.75	1.25
Lambda Decrement	0.999	0.0001	1.25
Weight Scale	0.01	0.0001	0.999
Momentum	0.5	0.0001	0.999
Error Ratio	1.02	1.0001	1.5

Source: [Rybchynski 2005].

The definition of these parameters were described in Ennett's paper [Ennett CM et al. 2004] :

"Learning rate (lr): The value of the learning rate determines the speed at which the network attains a minimum in the criterion function so long as it is small enough to insure convergence. If the learning rate is too high, it may oscillate around the global minimum, and is unable to converge.

Learning rate increment (lr_inc): The learning rate's incremental value.

Learning rate decrement (lr_dec): The learning rate's decrement value.

Weight-decay constant (λ): The weight-elimination constant determines how strongly the weights are penalised.

Weight-decay constant increment (λ_inc): The weight-decay constant's

incremental value.

Weight-decay constant decrement (λ_dec): The weight-decay constant's decrement value.

Weight-elimination scale factor (wo): Weight-elimination scale factor defines the sizes of "large" and "small" weights. When wo is small, the small weights will be forced to zero resulting in fewer large weights (i.e., weight-elimination). A large wo causes many small weights to remain and limits the size of large weights (i.e., weight-decay).

Momentum (momentum): The momentum parameter adds a proportion of the previous weight-change value to the new value, thereby giving the algorithm some "momentum" to prevent it from getting caught in local minima.

Error ratio (err_ratio): The error ratio controls how the backpropagation makes adaptive changes in the learning rate, the weight-decay constant, and the momentum term."

3.4 DT Model Creation

3.4.1 DT Model Created From See5

See5 is a software package for the creation and evaluation of DTs which can treat thousands of data in seconds. Therefore, in the first step, we used all of the independent variables ready for analysis mentioned in Table 3 as inputs for See5

analysis. All of the variables were used with their original values directly. The values of each variable did not need any treatment, which was different from what we did in LR models and ANN models.

A number of files including a names file, a data file, a test file and a cost file were needed to be created for See5 application. The names file was created manually in text format, in which each variable's name, type and numbers of values were specified. The data file was the training data created by SPSS and converted into comma delimited txt format.

The cost file was created by defining different misclassification costs, such as 0, 1: 2. This specifies that the cost of misclassifying positive cases (1) as negative cases (0) is 2 units. The unspecified error of misclassifying negative cases (0) as positive cases (1) has a default cost of 1 unit. In other words, the error of misclassifying positive cases as negative cases is two times more costly than the opposite misclassification. We can get a number of different classifiers by changing the value of cost. The performance of these classifiers can be plotted using a ROC curve. All of the See5 experiments applied default parameters for global pruning. The button of "Pruning CF" was set as 25% and the minimum cases were set as 2.

3.4.2 Input Variable Reduction

After the model construction with the entire set of 13 input variables, decisions were be made to eliminate unimportant inputs from the model. The variable selection was based on the altered performance of decision trees in response to changes in the input variable set. In this study, we tried to use the significant variables selected by

LR model as inputs to build another DT model to see if its performance on the test data could be improved.

3.4.3 Area Under the ROC Curve

As the ROC of See5 was plotted from a number of points, trapezoidal integration was used to calculate the AUC. Trapezoidal integration estimates the AUC by connecting all the points in ROC with straight lines instead of curves, which might lead to underestimation of the AUC. However, as long as there were enough points available for calculation, the area estimation is expected to be reasonable [Bradley AP, et al. 1997]. Furthermore, a trapezoidal approach was independent of any assumptions as to the underlying class distribution. Besides AUC, we still needed standard error of the AUC (SE(AUC)), which could be calculated from the standard error of the Wilcoxon statistic SE(W). The SE(AUC) is required so we could compare the significant difference of AUC among different classification approaches. In this study, we calculated area under AUC and SE(AUC) according to the following formula [Bradley AP, et al. 1997]:

$$AUC = \sum_i \{(1-\beta_i \Delta\alpha) + 0.5[\Delta(1-\beta)]\Delta\alpha \qquad (3.1)$$

where α = false positive rate=1-specificity

$1-\beta$ = sensitivity

$$SE(W) = \sqrt{\frac{\theta(1-\theta) + (C_p - 1)(Q_1 - \theta^{2)} + (C_n - 1)(Q_2 - \theta^2)}{C_p C_n}} \qquad (3.2)$$

Where θ is the area under ROC curve, $Q_1 = \theta/(2- \theta)$ and $Q_2 = 2\theta^2/(1+ \theta)$. C_p and C_n are the number of positive and negative cases respectively. Therefore, the 95% confidence interval of AUC is calculated using the following formula:

$$(AUC-1.96*SE(W),\ AUC+1.96*SE(W)) \qquad (3.3)$$

CHAPTER 4. RESULTS AND DISCUSSION

4.1 LR Models

4.1.1 Data Categorization

For most variables in Table 6, the differences in the prevalence of severe fall injury among categories were obvious. For example, the rate of severe fall injuries was higher in boys (7.88%) than in girls (6.66%). In the LR analysis, the category of girls was served as reference group, and the category of boys was compared with the category of girls. There were 5 categories in the variable of *age*. Since the rate of severe fall injury was the lowest in children of one year, this age group was used as the reference group. The other 4 age categories were expressed by dummy variables of $X_{2\text{-}1}$, $X_{2\text{-}2}$, $X_{2\text{-}3}$, and $X_{2\text{-}4}$ and were compared with the reference group.

For some variables like *Mf1*, the category with lowest rate was unknown factors, which was not suitable to be used as reference, so the second lowest category "stairs or steps" was selected as reference instead. Totally 28 dummy variables were used to represent the 13 original variables in the LR model creation shown in Table 6.

We applied the same data processing method to the test data set, which would be used to evaluate the final LR model performance.

4.1.2 Prediction Model

First, all of the dummy variables listed in Table 6 were used to establish a stepwise logistic regression model. Eight original variables (27 dummy variables in

total) were identified as significant variables for the model according to statistical test (OR values and their confidence intervals). Second, the non-significant variables were removed from the model. Only the 8 significant variables represented by 27 dummy variables were used to build up the final LR model. The coefficient of parameter, OR and its 95% confidence interval for each dummy variable in the final LR model were listed in Table 9. The estimate of interception was -6.936. The model creation was based on training dataset only.

For the variable of *sex*, the category of "girls" was used as reference group (shown in Table 6). The OR and its 95% confidence interval for "boys" were 1.165 (95% CI: 1.050-1.292). These results indicate that boys were significantly more likely to have a severe outcome of falls than girls as a result of a fall injury. This is consistent with the result in Table 7 that 7.88% of fall injuries in boys were severe, while only 6.66% fall injuries in girls were severe. This difference was significant based on LR model. Therefore, we concluded that *sex* was a significant prediction variable for the severity of falls for children. For the variable *locate*, the category "school" was used as a reference group, all of the other categories had higher prevalence rate of severe outcome and also the OR and 95% confidence interval for each dummy variable of variable locate were significant. Similarly, we can conclude that childhood falls which happened at school were significantly less likely to have a severe outcome compared with other locations.

In the case where not all dummy variables for one of the original variables were significant, but at least one dummy variable was significant, we considered the original variable to be significant. For example, for the variable *area*, OR values and

their 95% confidence intervals of its two dummy variables $X_{8\text{-}1}$ and $X_{8\text{-}4}$ were 1.159 (95% CI: 0.908-1.571) and 1.171 (95% CI: 0.995-1.435) respectively. Both of these were non-significant, which meant that there was no significant difference in the severity of childhood falls that happened at sports related areas ($X_{8\text{-}1}$), and other areas ($X_{8\text{-}4}$), compared with those at home rooms. However, childhood falls that happened at playground, garden, yard ($X_{8\text{-}2}$) and some unknown area ($X_{8\text{-}3}$), were more likely to have severe outcome than those that happened at home (reference category), based on statistical test for dummy variables $X_{8\text{-}2}$ and $X_{8\text{-}3}$. Their OR values and confidence intervals were 1.278 (1.021-1.600) and 1.400 (1.141-1.717), respectively. We therefore considered the variable of *area* as one of the significant predictors for the severity of childhood falls.

The β values in Table 9 were the coefficients of regression indicating the magnitude of change in the dependent variable (*disposition*) with corresponding change in the independent variable. The positive β values meant that this variable increased the probability of a severe outcome, while negative β values meant that this variable decreased the probability of severe outcome. Larger absolute β values meant that variable has a stronger effect on the probability of severe outcome.

The β value of variable $X_{1\text{-}1}$ was 0.152, which was positive. This meant that if sex was boys, the probability of severe outcome would be increased. The β values of variable $X_{8\text{-}1}$, $X_{8\text{-}2}$, $X_{8\text{-}3}$ and $X_{8\text{-}4}$ were 0.178, 0.246, 0.336 and 0.158 respectively. The reference group of the variable *area* was "Bedroom, living room, family room and recreation room". All of these four dummy variables had positive β values. In order to say if these dummy variables were significant, we needed to check OR values and

their 95% confidence intervals. As β values were always consistent with OR values and their 95% confidence intervals, we could see that the OR values of X_{8-2} with β value 0.246 and X_{8-3} with β value 0.336 were larger than that of X_{8-1} with β value 0.178 and X_{8-4} with β value 0.158. According to the 95% confidence intervals, the latter two β were not significant compared with reference category. Therefore, We could say that the probability of severe outcome would increase if fall injuries happened at "Playground, garden and yard" (X_{8-2}) or "Unknown area" (X_{8-3}) compared with those happened at house rooms (reference group).

As all of the LR models were created using the training data only, for the last step, the test data set was used to evaluate the model performance. Remember β value of each dummy variable was calculated from the training data (Table 8). For a given patient with values for each dummy variable, we could compute the likelihood (p value) of a severe outcome from equation 2.5. Therefore, the expected p values were calculated for test data. We could get different models by setting different cut-off values (p values). Table 10 describes the validity of the prediction models under different cut-off values. The model performance was similar between the training set and the test set according to the Table10. The sensitivity, specificity, and CCR of the prediction model were 66.47%, 85.54% and 67.86% respectively at the cut-off value of 0.06 for the test set, and 68.06%, 84.39% and 69.26% respectively at the same cut-off value for the training set. We considered that all of the three values (sensitivity, specificity and CCR) were reasonable at the cut-off value of 0.06.

Table 9 The association between severe outcome of fall injuries and significant predicting variables included in the logistic regression model

Variables	β[1]	OR[2]	95% CI[3]
Sex (X_{1-1})	0.152	1.165	1.050-1.292
Age (X_{2-1})	0.252	1.287	1.002-1.652
Age (X_{2-2})	0.148	1.159	0.933-1.441
Age (X_{2-3})	0.330	1.391	1.120-1.728
Age (X_{2-4})	0.221	1.248	1.000-1.562
Followup_flag (X_{6-1})	0.345	1.412	1.269-1.570
Locate (X_{7-1})	0.441	1.554	1.294-1.867
Locate (X_{7-2})	0.390	1.477	1.222-1.786
Locate (X_{7-3})	0.640	1.896	1.556-2.310
Locate (X_{7-4})	0.234	1.263	1.009-1.582
Area (X_{8-1})	0.178	1.159	0.908-1.571
Area (X_{8-2})	0.246	1.278	1.021-1.600
Area (X_{8-3})	0.336	1.400	1.141-1.717
Area (X_{8-4})	0.158	1.171	0.955-1.435
Cfl (X_{11-1})	-0.137	0.872	0.730-1.041
Cfl (X_{11-2})	0.301	1.351	1.103-1.654
Cfl (X_{11-3})	0.228	1.256	1.078-1.464
Nil (X_{12-1})	0.119	1.126	0.808-1.571
Nil (X_{12-2})	1.131	3.099	2.266-4.237
Nil (X_{12-3})	3.226	25.167	18.663-33.937
Nil (X_{12-4})	3.359	28.745	21.081-39.196
Nil (X_{12-5})	-0.921	0.398	0.251-0.631
Bpl (X_{13-1})	-0.708	0.493	0.342-0.710
Bpl (X_{13-2})	0.139	1.149	0.817-1.617
Bpl (X_{13-3})	2.133	8.442	5.846-12.191
Bpl (X_{13-4})	1.391	4.018	2.871-5.623
Bpl (X_{13-5})	2.838	17.083	12.086-24.144

[1] β: coefficient of regression indicating the magnitude of change in dependent variable (disposition) with corresponding change in independent variable.

[2] OR: odds ratio.

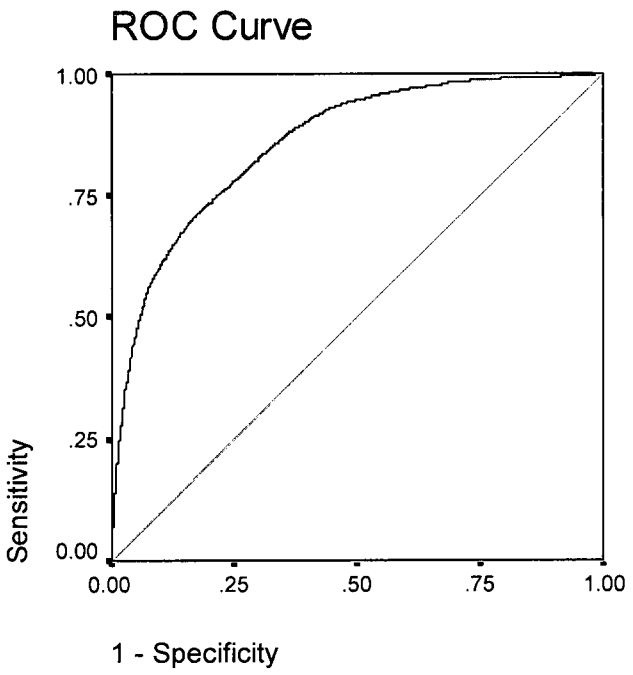
[3] CI: confidence interval.

Table 10 The validity of logistic predicting model under different cut-off values

Cut-off value	Training set			Test set		
	Se[1]	Sp[2]	CCR[3]	Se	Sp	CCR
0.01	36.80	97.22	41.25	36.81	95.90	41.14
0.02	42.98	96.22	46.90	42.66	94.76	46.48
0.03	49.91	94.54	53.20	49.23	93.85	52.50
0.04	55.50	92.86	58.26	54.31	92.48	57.11
0.05	62.12	89.08	64.11	60.56	89.64	62.70
0.06	68.06	84.39	69.26	66.47	85.54	67.86
0.08	78.83	74.71	78.53	77.63	75.40	77.47
0.10	85.82	67.39	84.47	85.64	67.54	84.31
0.15	92.18	56.90	89.58	92.29	56.49	89.66
0.20	93.54	52.59	90.52	93.85	51.82	90.77
0.30	96.41	39.80	92.23	96.21	39.41	92.05

[1]Se: sensitivity [2]Sp: specificity [3]CCR: correct classification rate

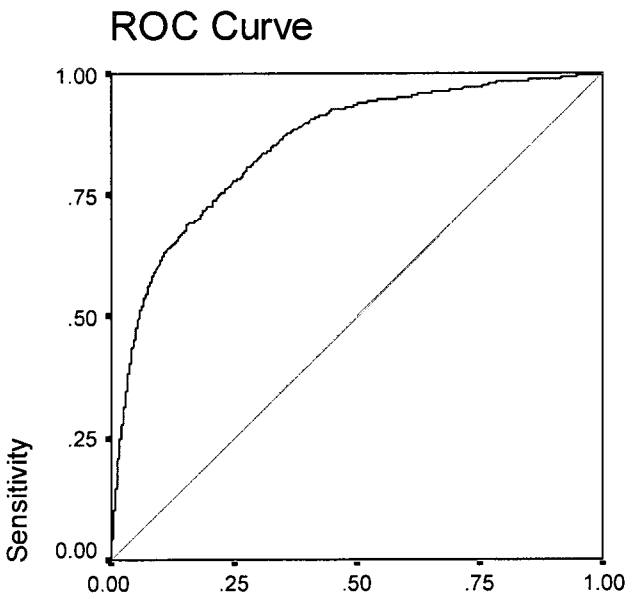
ROC curves were made by SPSS for the training and test sets separately (Figure 6 and Figure 7). SPSS automatically provided the AUC and 95% confidence interval of the AUC shown in below figures. The areas under ROC curves would be used to compare with DT and ANN models. The area under ROC curve was 0.859 (95% CI: 0.851-0.867) in the training set (Figure 6) and 0.852 (95% CI: 0.839-0.866) in test set (Figure 7).



Area Under the Curve

Area	Std. Error(a)	95% Confidence Interval	
		Lower Bound	Upper Bound
0.859	0.004	0.851	0.867

Figure 6 ROC curve and AUC for the training set in LR analysis



Area Under the Curve

Area	Std. Error(a)	95% Confidence Interval	
		Lower Bound	Upper Bound
0.852	0.007	0.839	0.866

Figure 7 ROC curve and AUC for the test set in LR analysis

4.2 DT Models

4.2.1 DT Models Created by Raw Variables

The performance of DT models created with See5 using 14 raw variables (listed in Table 3 including *ID*) as inputs is shown below. The variable *ID* was just a label, which was not used by See5 for classification purpose. For the other 13 input variables, the original values were used to build DT models. The outcome of *disposition* was binary with values of 1 indicating severe outcome and 0 indicating non-severe outcome of a fall injury. See5 automatically applied the training data to build the classifier and the test data to test the performance of the classifier. The performance of decision tree models from See5 with different misclassification costs for both training data and test data is summarized in Table 11.

In Table 11, tree size is the number of non-empty leaves on the tree. By changing the misclassification costs, we got a number of DT models with different tree sizes, error rates, sensitivities and specificities. The sensitivity and error rate were increasing and the specificity was decreasing when the cost of misclassifying positive cases (1) as negative cases (0) was increasing. When the cost of misclassifying positive cases (1) as negative cases (0) was 12, the sensitivity was 69.6 and the specificity was 82.5 for test set. Although the sensitivity could be improved further when the even higher cost values were used, the specificity would decrease and the error rate would increase dramatically. Therefore, we considered that the misclassification cost of "0,1:12" was the optimum value to create DT models for this research purpose.

Table 11 See5 performance using entire input variables with different misclassification costs on training set and test set

Misclassification cost	Tree size	Error rate (%)		Sensitivity (%)		Specificity (%)	
		Tr[1]	Te[2]	Tr	Te	Tr	Te
0,1: 1	4	6.4	6.4	18.7	18.7	99.5	99.5
0,1: 2	12	7.0	7.5	32.0	30.5	97.9	97.4
0,1: 4	27	10.0	10.2	58.3	53.9	92.5	92.6
0,1: 12	27	18.1	18.5	71.3	69.6	82.7	82.5
0,1: 16	20	32.9	33.5	87.9	86.6	65.4	64.9
0,1: 30	31	40.3	41.0	95.2	91.7	56.8	56.4
0,1: 60	20	51.0	51.5	97.1	94.2	45.2	44.9

[1]Tr: Training [2]Te: Test

The confusion matrix is another way to show the classification results. Table 12 described the classification results for training set by confusion matrix for the best DT model when the misclassification cost was "0,1: 12". Table 13 described the same thing for test set. The CCR of training and test sets were 81.9% and 81.5% respectively, which were consistent with the error rates shown in table 11.

Table 112 Confusion matrix of the best DT model for training set

		Predicted class		
		Yes	**No**	**Total**
True class	**Yes**	1,489	599	2,088
	No	4,537	21,691	26,228
	Total	6,026	22,290	28,316

CCR = (1489+21691)/28316 = 81.9%

Table 13 Confusion matrix of the best DT model for test set

		Predicted class		
		Yes	**No**	**Total**
True class	**Yes**	611	267	878
	No	1,942	9,149	11,091
	Total	2,553	9,416	11,969

CCR = (611 + 9149) / 11969 = 81.5%

To clearly illustrate the model performance over the entire range of sensitivity and specificity, ROC curves were plotted for both training and test sets manually according to the results in Table 11. In Figure 8, the points of setting misclassification cost as "0,1:12" are the closest ones to the top-left corner of the ROC curves for both training and test sets, which represent the best DT model.

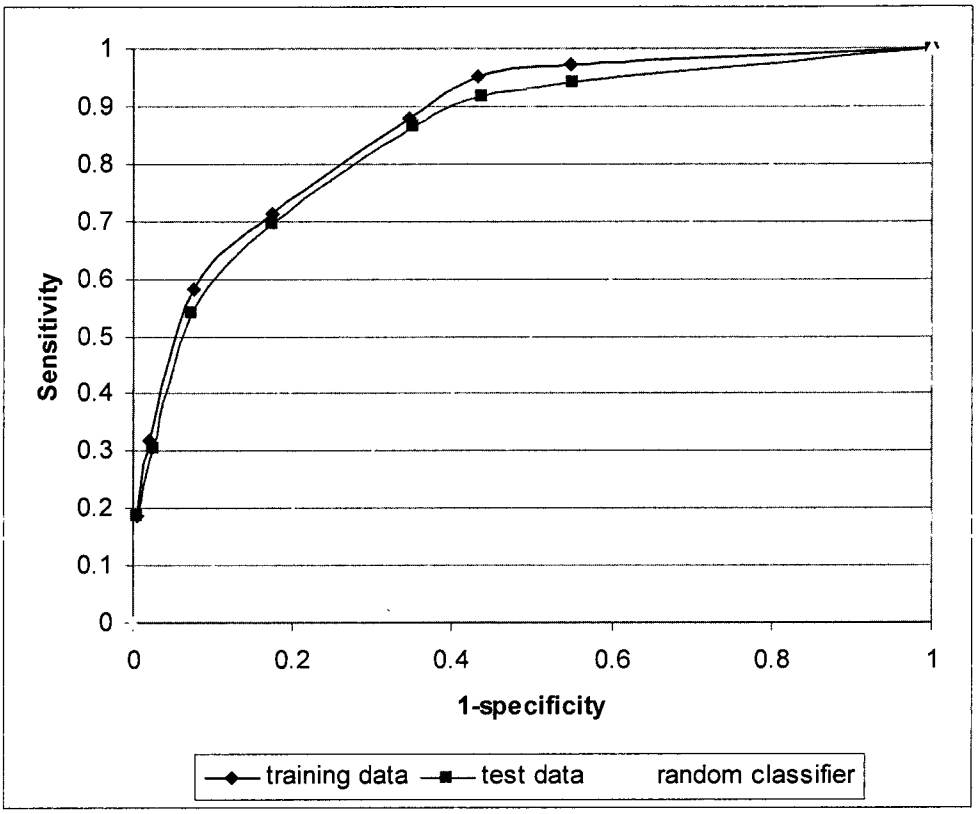


Figure 8 ROC curves for training set and test set in DT models using entire raw inputs

According to the equations of 3.1, 3.2 and 3.3, the area under the ROC curve was 0.866 with standard error of 0.005 for training set, and 0.844 with standard error of 0.008 for test set. The area under the curve and its 95% confidence intervals were 0.866 (95% CI: 0.856-0.876) and 0.844 (95% CI: 0.828-0.860) for training set and test set, respectively. The results were very close to LR models, though the AUC of test set was a little bit lower than that in the final LR model.

4.2.2 DT Models Created by Reduced Variables

The performance of See5 using reduced variables selected by logistic regression is summarized below. From LR models, 8 variables were identified to be significantly associated with the severe outcome of fall injuries (Table 9). For DT model creation, the original values of those 8 variables were used, instead of the dummy variables.

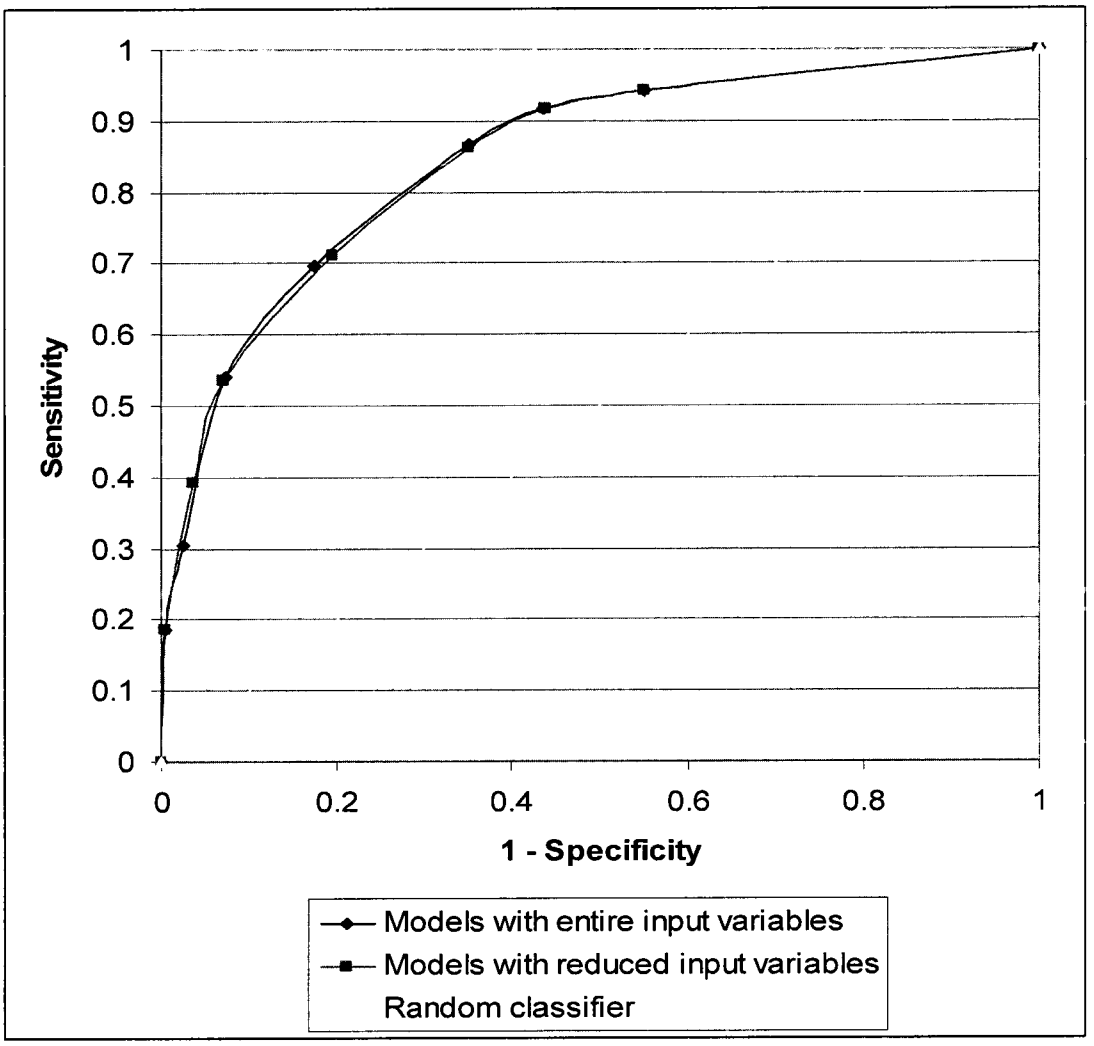
Table 14 See5 performance using reduced input variables with different misclassification costs on training data and test data.

Misclassification cost	Tree size	Error rate (%)		Sensitivity (%)		Specificity (%)	
		Tr[1]	Te[2]	Tr	Te	Tr	Te
0,1: 1	4	6.4	6.4	18.7	18.7	99.5	99.5
0,1: 2	117	6.9	7.8	45.2	39.3	97.0	96.4
0,1: 4	27	9.9	10.0	57.5	53.4	92.7	92.9
0,1: 12	27	19.9	20.3	73.1	71.1	80.7	80.4
0,1: 16	25	32.8	33.5	88.1	86.1	65.6	65.0
0,1: 30	32	40.5	41.2	95.4	91.7	56.6	56.2
0,1: 60	19	51.0	51.5	97.1	94.2	45.2	44.9

[1]Tr: Training; [2]Te: Test

The performance was similar for models trained using all 14 original input variables and models trained using only the LR-selected reduced set of eight significant variables, based on error rate, sensitivity and specificity. ROC curves were plotted on the two test sets to compare the performance between these two models.

Figure 9 provides a summary of the performance measure of decision tree models trained using all input variables and trained using the reduced set of input variables.



Area Under the Curve

	AUC	SE(AUC)	95% CI
DT models with 13 input variables	0.844	0.008	0.829-0.860
DT Models with 8 input variables	0.844	0.008	0.829-0.860

Note: The calculations were based on equations of 3.1, 3.2 and 3.3.

Figure 9 ROC curves and AUC for test sets in DT models using entire inputs and reduced inputs

The AUC and its 95% confidence intervals of DT models based on reduced input variables and entire input variables were exactly the same (0.844, 95% CI: 0.829-0.860), which meant that using reduced variables as inputs could produce the same performance as using entire input variables. This could be explained by the reason that DTs apply filter approach as an embedded feature selection method in the tree building process and this approach worked well for this research purpose using CHIRPP data.

Furthermore, See5 provided the relative contribution of each input variable to the trained models. Variable *Nil* (Nature of injury suffered by the injured persons) and *BPI* (injured body part) are always the most contributive variables in both created DT models (i.e. models trained using full set of 14 variables, or models trained using reduced set of 8 variables).

4.3 ANN Models

4.3.1 Data Pre-processing

The raw data were randomly split into training, test and validation data sets. From the ten subsets of validation set, average evaluation of the performance of the models was calculated. Numbers of cases in each set were shown in Table 7. There were 22,894 cases in training set, 8,376 cases in test set and 11,969 cases in validation data sets available for ANN model creation. The severe outcome cases in the training set were re-sampled to comprise of 23.9% of cases.

4.3.2 Input Variables

Preliminary results showed that using dummy variables as ANN inputs could improve the performance of ANN models dramatically for this research purpose, so all of the dummy variables listed in Table 6 were used as ANN inputs for the first step (results not shown). To reduce the ANN model complexity and improve the model performance further, only the significant variables selected by LR analysis were used as ANN inputs for the final ANN model creation. A Table 13 showed the eight variables (24 variables in total, because some of them are dummy variables) determined to be significant following LR analysis and used for the final ANN model creation. Most of them were dummy variables except for *age* (continuous), *sex* (binary) and *followup_flag* (binary).

Table 15 Variable list for ANN inputs

Variables	Variable type
Sex (X_1)	Binary
Age (X_2)	Continuous
Followup_flag (X_3)	Binary
Locate (X_{7-1})	Dummy
Locate (X_{7-2})	Dummy
Locate (X_{7-3})	Dummy
Locate (X_{7-4})	Dummy
Area (X_{8-1})	Dummy
Area (X_{8-2})	Dummy
Area (X_{8-3})	Dummy
Area (X_{8-4})	Dummy
Cf1 (X_{11-1})	Dummy
Cf1 (X_{11-2})	Dummy
Cf1 (X_{11-3})	Dummy
Ni1 (X_{12-1})	Dummy
Ni1 (X_{12-2})	Dummy
Ni1 (X_{12-3})	Dummy
Ni1 (X_{12-4})	Dummy
Ni1 (X_{12-5})	Dummy
Bp1 (X_{13-1})	Dummy
Bp1 (X_{13-2})	Dummy
Bp1 (X_{13-3})	Dummy
Bp1 (X_{13-4})	Dummy
Bp1 (X_{13-5})	Dummy

4.3.3 Best ANN Model Performance

The ANN with one hidden layer and weight-elimination cost function was trained and tested for the best performance based on sensitivity and specificity using the log-sensitivity index as stopping criteria. The best performance of the ANN model with 5 hidden nodes predicting the severity of childhood fall injuries was shown in Table 14. From 10 validation data sets the average performance of the ANN model on the validation set could be calculated. The average AUC calculated from validation sets was 0.844 (95% confidence intervals: 0.825-0.863). This was almost the same as the AUC from DT models on test data and a little bit lower than the AUC from the final LR model on test data. All of the values of AUC from these three models (LR, DTs and ANNs) can be considered as good.

Table 16 The best performance of the ANN model predicting the severity of childhood fall injuries

	Se (%)	Sp (%)	CCR	Log_se Index	AUC
Training	61.5	90.8	85.1	0.2180	0.8644
Test	62.4	90.1	88.0	0.2234	0.8608
Validation average	59.2	90.6	88.3	0.1980	0.8440

Se: sensitivity Sp: specificity CCR: correct classification rate
Log_sen Index: Log_sensitivity Index
AUC: area under receiver operating characteristic curve

The following graphs of ROC curves for the optimal performance of the ANN in the prediction of childhood fall injuries were created by the ANN tool automatically (Figure 10).

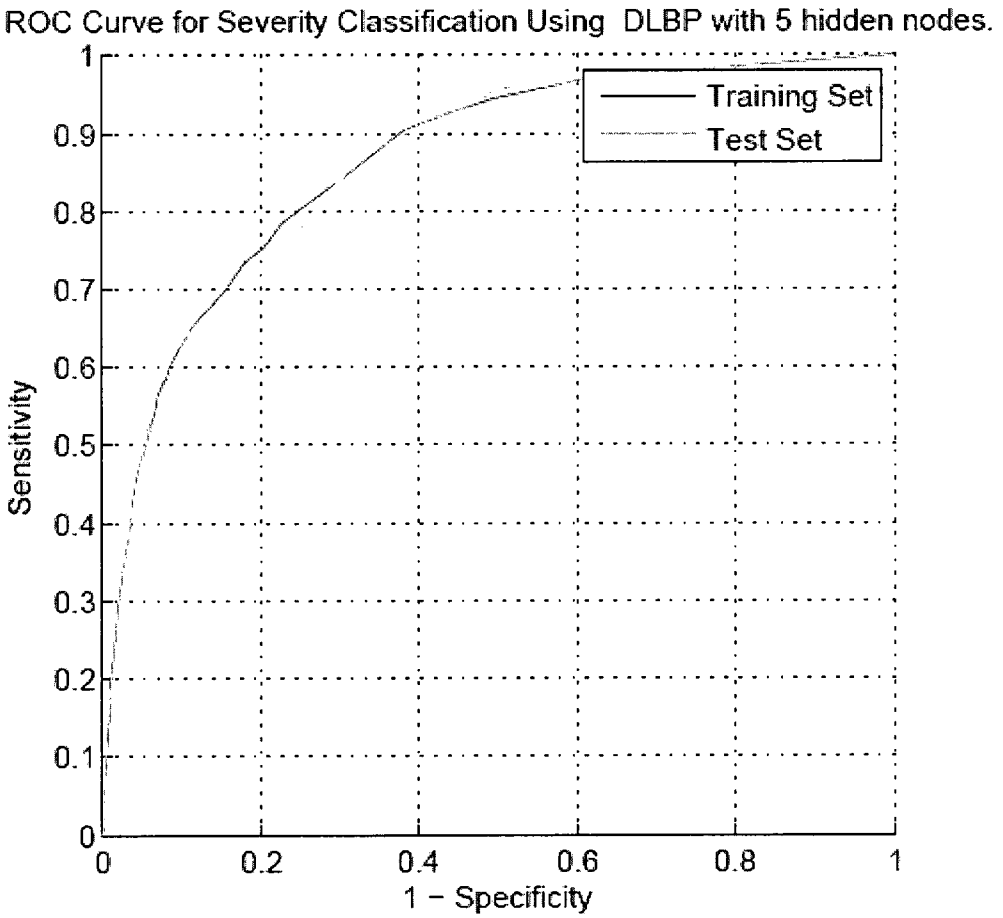


Figure 10 ROC curve for the serevity classification using DLBP with 5 hidden nodes

4.4 Comparison of the performance of LR, DT and ANN models

In order to compare the performance of the three models on the severity of childhood fall injury prediction, the AUC, sensitivity, specificity and CCR of LR, DT and ANN models were listed in Table 16.

All of these three models created very similar results according to the values of AUC. LR is a little better than DTs and ANNs in term of AUC. However at the sensitivity around 60% and specificity around 90%, both DT and ANN models had higher CCR compared with LR model. Because we had far more negative cases than positive cases in the dataset and the specificities of DTs and ANNs were higher than that of LR in Table16, this might be the reason why the CCR of LR was lower than those of DT and ANN. We selected these specific values to compare these three models, because the ANN tool provided the values of the best performance on validation sets at this point.

Table 17 Comparison of LR, DT and ANN in childhood fall injury severity prediction analysis

	AUC	Sensitivity (%)	Specificity (%)	CCR
LR Model	0.852 (0.839 - 0.866)	60.6	89.6	62.7
DT Model	0.844 (0.829 - 0.860)	58.0	90.7	88.7
ANN Model	0.844 (0.825 – 0.863)	59.2	90.6	88.3

In this study, both LR and ANN model creation required more work in data preparation than DT model creation. For example, we needed to combine the values

for most of the nominal variables and create dummy variables, which could be used for LR and ANN model creation. We needed to develop SAS programs to create LR models and made ROC curves from SPSS manually. Although the training of ANNs was time- consuming, the ANN tool developed by MIRG was very convenient and robust and almost all of the results and figures were created automatically. For DTs, no data preparation was needed. The training and testing was very fast. We could get the model performance over the entire range of sensitivity and specificity by only changing the values in the cost file. Therefore, it is a very useful tool for this study.

4.5 Discussion

In this study, LR, DT and ANN models were developed to analyze the severity of childhood fall injuries in 2007 based on the CHIRPP dataset and to assess the feasibility of predicting severe childhood fall injuries. All three techniques showed high predictive power on this study issue, which suggested that these models might have potential clinical usefulness for determining prognosis and appropriate treatment. The identified important factors by the developed models can be used for fall injury prevention as discussed further below.

4.5.1 Applying Logistic Regression, Decision Trees and Artificial Neural Networks to Predict Severe Childhood Fall Injuries Using CHIRPP Data

CHIRPP is a nationwide dataset, which collected detailed information on childhood injuries. This study focused on fall injuries since they are very common for children and sometimes may cause severe consequences. The dataset available for us provides an opportunity to compare traditional LR model with DTs and ANNs. LR

models are traditionally used for medical binary outcome prediction and data mining approaches, whereas DTs and ANNs require few assumptions about the data distribution and thus maybe more appropriate for noisy data sets like CHIRPP.

According to the results of this study, LR, DTs and ANNs had nearly identical performance for the prediction of severe childhood fall injuries measured by sensitivity, specificity and the AUC. LR is a little better than DTs and ANNs according to AUC, which is a measure of discrimination of predictive models.

From the physicians' point of view, the specificity should be about 85% for the predictive models. Even lower specificity or higher sensitivity would cause more negative (non-severe) cases to be admitted to the hospital, which would increase the cost of hospitalization. Based on this criteria, we selected the best LR model at the cut off p value of 0.06 (Table 10) and the best DT model at the misclassification cost of 0,1: 12 (Table 11), which means the cost of misclassifying positive cases (represent by 1) as negative cases (represent by 0) was 12. . In other words, at this point the error of misclassifying positive cases as negative cases is four times more costly than the opposite misclassification.

In this study, compared with LR and DTs, although ANNs reached almost the same sensitivity, specificity and AUC, the training process of ANNs was highly time consuming, especially when dummy variables were introduced which caused the complexity of the ANN models to increase significantly.

Compared with ANNs, decision trees can be viewed as a white box, which is easy to be interpreted by a human expert about the relation between inputs and outputs. In this study, we benefited from the following characteristics of DT models.

First, it can handle both numerical and categorical data directly requiring little data preparation. All variables from CHIRPP could be used directly no matter the variables were continuous or nominal. Second, decision trees can handle datasets that may have errors or missing values. The built-in feature selection method works well for this study. Third, the DT models are typically robust and can handle large amounts of data in a relatively short time. The process of DT model creation is very fast and takes only a few minutes for large datasets such as those used in the present study. Finally, the performance of DT models created from See5 was as good as ANNs and LR.

However, health professionals always considered that the traditional LR models are more reliable than machine learning tools. Although DT models were considered as the most convenient tool for this study due to exceptional merits discussed above, health professionals might not be comfortable with this data mining tool unless the results of DTs performed better than LR.

4.5.2 Applying Logistic Regression as feature selection for DTs and ANNs

In this study, the LR was used as feature selection method for both DT and ANN models. LR models were built in the first step, from which 8 significant variables – *Sex, Age, Followup-flag, Locate, Area, Cf1, Ni1* and *Bp1* (Table 9) were identified. Then these variables were used as inputs to build DT and ANN models.

The best performance of ANN models could be reached by using the 8 significant variables selected by LR. However, DT models produced exactly the same results by using the 8 significant variables as using entire 13 variables, which meant that See5 could effectively take advantage of useful variables regardless the existence

of less useful variables. It seemed that using LR as feature selection method did not work well for DTs, but may work for ANNs for this study.

4.5.3 Important Variable Identification for Childhood Fall Injuries

LR model is a generalization of linear regression to gain insight into the relative contribution of the independent variables to the outcome. From LR analysis, 8 significant variables were identified (Table 9). Among them, variable *Nil* and *Bpl* had bigger OR values compared with other variables, which meant that the magnitude of the association between *Nil*, *Bpl* and the severity outcome was stronger than other variables. Similarly, these two variables were always considered as the most contributive variables in creating DT models. Therefore, among the eight significant variables, we concluded that variables of *Nil* and *Bpl* are the most two important variables for the prediction of the severity of childhood fall injury.

The variable of *Nil* indicates the nature of injury suffered by the injured person. We defined "Open wound" as reference group for this variable. The other three groups of "Minor head injury", "Fracture" and "Dislocation; injury to nerve, blood vessel or internal organ; multiple injuries, etc" were significantly more likely to have severe outcome than the reference group. The OR values and 95% confidence intervals reached 3.10 (2.27-4.24), 25.17 (18.66-33.94) and 28.75 (21.08-39.20) for the three groups respectively.

The variable of *Bpl* indicates the injured body part. The body part of "Face and foot" was defined as reference group. It is significantly more likely to have severe outcome if the injured body parts were "Forearm; lower leg" or "Upper arm; elbow". The OR values and 95% confidence intervals reached 8.44 (5.85-12.19) and 4.02

(2.87-5.62) for the two groups, respectively. Therefore, detail information about these two variables would further improve the LR and DT model performance.

Knowing the contributing factors to severe fall injuries can help inform future injury severity reduction programs. For example, as even minor head injury could cause severe outcome, wearing helmet when biking was legislated. Fracture could cause severe outcome as well, so health professionals encouraged children to wear elbow pads and knee pads in some sports like skating. Based on these two characteristics of fall injury, the nurses in triage could correctly predict the treatment of the injured children.

The most important variables from ANN models were not identified in this study. However, a method coined as the Garson-Goh approach and developed by MIRG researchers can be used to determine the weights of each input in a three-layered ANN [Rybchynski 2005]. Due to time constraints, this has been left as future work.

CHAPTER 5. CONCLUSIONS and FUTURE WORK

5.1 Conclusions

Three models - LR, DT and ANN were developed in this study to predict severe childhood fall injuries using CHIRPP database. We defined severe childhood fall injuries as those needed prolonged observation in emergency rooms or admission to a hospital. The outcome was represented by a binary categorical variable, in which severe fall injuries were represented with a value of "1" and non-severe fall injuries were represented with a value of "0". In order to achieve an unbiased measure of accuracy and to fairly compare the performance of the three methods; we used a test dataset which was never used in the process of model creation to evaluate the performance of the models.

All models showed strong prediction ability with promising sensitivity and specificity levels. Using LR models to select significant variables as inputs for ANN allowed us to reduce the number of inputs for ANNs to get more reasonable results. While ANNs were able to obtain similar accuracy to DT models, the most significant shortcoming of ANN is that the training process is time consuming. See5 also appears to be an efficient tool for the construction of DT models for this study.

The research results showed that the severity of childhood fall injuries can be predicted by both statistical and data mining tools by using CHIRPP data. We also identified two important variables – *Ni1* and *Bp1* from both LR and DT models.

5.2 Future Work

Childhood falls comprise a leading cause of injury for emergency room attendance in most industrialized countries. Both statistical method and data mining tools appear to be efficient approaches for severity prediction. However, there were still some issues that need to be considered for future studies in this subject.

1) Due to the large sample size and several dummy variables created for CHIRPP data, the ANN tool needs to be further improved to reduce the training time.

2) The most contributive variables from ANNs should be identified in future work.

3) Since we only had one year (2007) CHIRPP data, the finding from this study may be limited. Extending the study to multiple years may provide more generalizable results.

4) Collecting more detail information on the two important variables (*Ni1* and *Bp1*) identified by LR and DT models was expected to improve the performance of these models further.

REFERENCES

[Allison PD 1999]
Allison PD 1999 Logistic Regression Using SAS: Theory and Application. SAS Institute Inc. Cary, NC, USA. 1999.

[Angus DE et al. 1998]
Angus DE, Cloutier E, Albert T, Chenard D, Shariatmadar A, et al. The Economic Burden of Unintentional Injury in Canada. SMARTRISK. 1998

[Begg RK et al. 2006]
Begg RK, Kamruzzaman J, Sarker RA. Neural network in Healthcare: Potential and Chanllenges. IGI Global 2006. p

[Bradley AP et al 1997]
Bradley AP. The Use of Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms. Pattern Recognition. 1997; 30(7): 1145-1159.

[Bishop 1995]
Bishop C. Neural Networks for Pattern Recognition. 1995. Oxford: Clarendon Press.482p.

[Boer SD et al 2005]
Sylvia SB, Keizer NF, Jonge ED. Performance of Prognostic Models in Critically Ill Cancer Patients – a Review. Crit Care. 2005; 9(4): R458-463.

[Child and Youth Injury in Review 2009]
Child and Youth Injury in Review 2009. Public Health Agency of Canada. http://www.phac-aspc.gc.ca/publicat/cyi-bej/2009/index-eng.php

[Chizi B et al 2002]
Chizi, B. and Maimon, O. "On Dimensionality Reduction of High Dimensional Data Sets" In Frontiers in Artificial Intelligence and Applications. IOS press, 2002: 230–236p.

[Christersen 1997]
Christersen R. Log-Linear Models & Logistic Regression. New York, NY, USA: Springer-Verlag New York, Incorporated, 1997.

[Clermont G et al. 2001]
Clermont G, Angus D, DiRusso S, Griffin M, Linde-Zwirble W. Predicting Hospital Mortality for Patients in the Intensive Care Unit: a Comparison of Artificial Neural Networks with Logistic Regression Models. Crit Care Med 2001;29:291-6.

[Committee on Injury and Poison Prevention 2001]
Falls From Heights: Windows, Roofs, and Balconies. Pediatrics 2001; vol.107 No.5 May: 1188-1191.

[Cook et al. 2000]
Cook D, Dixon P, Duckworth WM, Kaiser MS, Koehler K, Meeker WQ, and Stephenson WR. Chapter 3: Binary Response and Logistic Regression Analysis, in Beyond Traditional Statistical Methods. February 2001.
www.public.iastate.edu/~stat415/stephenson/stat415_chapter3.pdf

[Crawley T 1996]
Teri Crawley. Childhood Injury: Significance and Prevention Strategies. J Pediatr Nurs. 1996; 11(4) : 225-232.
[Delen D et al. 2005]
Delen D, Walker G, Kadam A. Predicting Breast Cancer Survivability: a Comparison of Three Data Mining Methods. Artif Intell Med 2005;34:113-27.
[D'Agostino RB et al. 2006]
D'Agostino RB, Sullivan LM, Beiser AS. Introductory Applied Biostatistics. Thomson. USA. 2006. Chapter 11.
[Dietz et al. 2002]
Dietz K, Gail M, and Krickeberg K. Logistic Regression: A Self-Learning Text (2nd Edition). New York, NY, USA: Springer-Verlag New York, Incorporated, 2002.
[DiRusso SM et al. 2002]
DiRusso SM., Chahine AA., Sullivan T, et al. Development of a Model for Prediction of Survival in Pediatric Trauma Patients: Comparison of Artificial Neural Networks and Logistic Regression. J.Pediatr.Surg. 2002;37(7):1098-1104.
[Dreiseitl S et al. 2002]
Dreiseitl S, Ohno-Machado L. Logistic Regression and Artificial Neural Network Classification Models: a Methodology Review. J.Biomed.Inform. 2002;35(5- 6):352-9.
[Dreyfus G 2005]
Dreyfus G. Neural Networks : Methodology and Applications. Heidelberg, DEU: Springer-Verlag, 2005. p5.
[Dybowski R et al. 1996]
Dybowski R, Gant V, Weller P, Chang R. Prediction of Outcome in the Critically Ill Using an Artificial Neural Network. The Lancet 1996;347(9009):1146-50.
[Eftekhar B et al. 2005]
Eftekhar B, Mohammad K, Ardebili HE, Ghodsi M, Ketabchi E. Comparison of Artificial Neural Network and Logistic Regression Models for Prediction of Mortality in Head Trauma Based on Initial Clinical Data. BMC.Med.Inform.Decis.Mak. 2005;5:3
[Ennett CM et al. 2000]
Ennett CM, Frize M. Selective Sampling to Overcome Skewed a Priori Probabilities with Neural Networks. Proceeding of the A.M.I.A. (American Medical Informatics Association) Annual Symposium. 2000: 225-229.
[Ennett CM et al. 2002]
Ennett CM, Frize M, Scales N. Logarithmic-sensitivity Index as a Stopping Criterion for Automated Neural Networks. EMBS/BEES Conference, 2002 (l): 74-75.
[Ennett CM 2003]
Imputation of Missing Values by Integrating Artificial Neural Networks and Case-based Reasoning. PhD in Electrical Engineering Thesis,

Carleton University, Ottawa, ON, Canada. 2003.

[Ennett CM et al. 2003]

Ennett CM, Frize M. Weight-elimination Neural Networks Applied to Coronary Surgery Mortality Prediction. IEEE Transactions of Information Technologies in Biomedicine. 2003; vol 7 (2): 86-92.

[Ennett CM et al. 2003 b]

Ennett CM, Frize M. Weight-elimination Neural Networks for Mortality Prediction in Coronary Artery Surgery. IEEE Trans Info Technol Biomed 2003;7(2):86–92.

[Ennett CM et al. 2004]

Ennett CM, Frize M, Charette E. Improvement and Automation of Artificial Neural Networks to Estimate Medical Outcomes. Medical Engineering and Physics. 2004; vol 26 (4): 321-328.

[Erdebil Y et al. 2005]

Erdebil Y, Frize M. An Analysis of CHIRPP Data to Predict Severe ATV Injuries Using Artificial Neural Networks. IEEE/EMBS Conf. Shanghai, 2005, Sept: 871-874.

[Fauske KM 2006]

Fausker KM. Example: Neural Network. 2006 http://www.texample.net/kitz/examples/neural-network.

[Public Health Agency of Canada on-line]

Public Health Agency of Canada. Facts on Injury (online). http://www.phac-aspc.gc.ca/publicat/lcd-pcd97/Table1-eng.php

[Fawcett T 2006]

Fawcett T. An Introduction to ROC Analysis. Pattern Recognition Letters. 2006; 27: 861-874.

[Flavin MP et al. 2006]

Flavin MP, Dostaler SM, Simpson K, Brison R, Pickett W. Stages of Development and Injury Patterns in the Early Years: a Population-Based Analysis. BMC Public Health. 2006,6: 187-197.

[Freund Y et al. 1999]

Freund Y, Schapire RE. A Short Introduction to Boosting. Journal of Japanese Society for Artificial Intelligence. 1999;14:771–780.

[Frize M et al. 1995]

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 et al. 2000]

Frize M, Ennett CM, Charette E. Automated Optimization of the Performance of Artificial Neural Networks to Estimate Medical Outcomes. Proc IEEE ITAB-ITIS 2000: 168-173.

[Fu 1994]

Fu L. Neural Networks in Computer Intelligence. United States of America, McGrawHill. 1994.

[Garson GD. 2009]

Garson GD. Logistic Regression.
http://faculty.chass.ncsu.edu/garson/PA765/logistic.htm

[Ge G et al 2008]
Ge G, Wong GW. Classification of Premalignant Pancreatic Cancer Mass-Spectrometry Data Using Decision Tree Ensembles. BMC Bioinformatics 2008, 9: 275.

[Gordon RA 2010]
Gordon RA. Regression analysis for the social sciences. Routledge. 2010 Feb.

[Han J et al 2001]
Han J, Kamber M. Data Mining: Concepts and Techniques. Morgan aufmann Publishers. 2001.

[Hand DJ 1981]
Hand DJ. Discrimination and Classification. Chichester: Wiley; 1981.

[Hecht-Nielsen 1987]
Hecht-Nielsen. Kolmogorov's Mapping Neural Network Existence Theorem. Proc. First IEEE Int. Joint Conf. Neural Netw. 111(1987) p11-14.

[Hinton GE et al. 1992]
Hinton GE. How Neural Networks Learn from Experience. Sci Am 1992;267(3):144-51.

[Hosmer D et al. 2000]
Hosmer D; Lemeshow S. Applied Logistic Regression. New York: wiley; 2000.

[Injury surveillance On-line]
Injuy surveillance on-line. Injury Mortality/Hospital Separations Charts.

[Jaimes F et al. 2005]
Jaimes F, Farbiarz J , lvarez D, rtinez C. Comparison Between Logistic Regression and Neural Networks to Predict Death in Patients with Suspected Sepsis in the Mergency Room. Crit Care 2005;9(2):R150-R156

[Larsen. 2008]
Larsen PV. Chapter 14: Logistic Regression, in Master of Applied Statistics. February 2008 http://statmaster.sdu.dk/courses/st111

[Leondes CT et al. 2003]
Leondes CT et al. Intelligent Systems: Technology and Applications. Volume II: Fuzzy Systems, Neural Networks, and Expert Systems CRC Press LLC 2003. Chapter1: Neural Network Techniques and Their Engineering Applications.

[Li YC et al. 2000]
Li YC, Liu L, Chiu WT, Jian WS. Neural Network Modeling for Surgical Decisions on Traumatic Brain Injury Patients. Int J Med Inform 2000;57:1-9.

[Liu H et al. 1998]
Liu H, Motoda H. Feature Selection for Knowledge Discovery and Data Mining. Boston: Kluwer Academic Plublishers. 1998. p33-p35.

[Mackey M et al. 2006]
Child Safety Good Practice Guide: Good Investments in Unintentional Child

Injury Prevention and Safety Promotion. Amesterdam, European Child Safety Alliance (Eurosafe), 2006.

[Mackenzie SG et al. 1999]
Susan G Mackenzie, Ivan Barry Pless. CHIRPP: Canadian's Principal Injury Surveillance Program. Injury Prevention 1999; 5: 208-213.

[Maimon O et al 2005]
Maimon O, Rokach L. Data Mining and Knowledge Discovery Handbook. New York: Springer. 2005.

[Menard SW 2002]
Menard SW. Applied Logistic Regression Analysis. Second edition . Sage Publication, Inc. 2002. Chapter 3.

[Nguyen T et al. 2002]
Nguyen T, Malley R, Inkelis SH, Kuppermann N. Comparison of Prediction Models for Adverse Outcome in Pediatric Meningococcal Disease Using Artificial Neural Network and Logistic Regression Analyses. J.Clin.Epidemiol. 2002;55(7):687-95.

[Pagano M et al. 2000]
Pagano M, Gauvreau K. Principles of Biostatistics, Second Edition. Pacific Grove, CA, USA: Duxbury Thomson learning; 2000.

[Paliwal M et al. 2009]
Paliwal M, Kumar UA. Neural Networks and Statistical Techniques: A Review of Applications. Expert System with Applications 2009;36:2-17.

[Parks RW et al. 1998]
Parks RW, Levine DS, Long DL. Fundamentals of Neural Network Modeling: Chap1- An introduction to Neural Network Modeling: Merits, Limitations, and Controversies. 1998 The MIT Press.

[Perlich C et al. 2003]
Perlich C, Provost F, Simonoff JS. Tree Induction vs. Logistic Regression: a Learningcurve Analysis. J Mach Learn Res 2003;4:211-55.

[Perner P et al. 2004]
Petra Perner, Chid Apte. Empirical Evaluation of Feature Subset Selection Based on a Real-World Data Set. Engineering Applications of Artificial Intelligence. 2004; 17: 285-288.

[Pickett W et al. 2003]
Pickett W, Streight S, Simpson K, Brison RJ. Injuries Experienced by Infant Children: a Population-Based Epidemiological Analysis. Pediatrics. 2003; Apr;111(4 Pt 1):e365-70.

[Public Health agency of Canada online]
Public Health agency of Canada. Facts on Injury (online), http://www.phac-aspc.gc.ca/publicat/lcd-pcd97/Table1-eng.php

[Qi L 2005]
Network Tool for Neonatal Intensive Care Units. Master of Science in Information and Systems Science Thesis. Carleton University, Ottawa, ON, Canada. May 2005.

[Quinlan JR 1986]
Quinlan JR. Induction of Decision Trees. Machine Learning. 1986; 1(1): 81–

106.
[Quinlan JR 1987]
Quinlan JR. Simplifying Decision Trees. International Journal of Man-Machine Studies. 1987; 27: 221–234.
[Quinlan JR 1988]
Quinlan JR. Decision Trees and Multivalued Attributes. Machine Intelligence. 1988; Oxford, England, Oxford Univ. Press, V. 11, pp. 305–318.
[Quinlan JR 1993]
Quinlan JR. C4.5: Programs for Machine Learning. Morgan Kaufmann. 1993.
[Quinlan JR 1996]
Quinlan JR. Bagging, boosting and C4.5. In Thirteenth National Conference on Artificial Intelligence. Portland 1996; 725–730.
[Raymcbride Online 2010]
Raymcbride. Artificial Neural Networks. 2010; Janurary. http://raymcbride.com/2010/01/15/artificial-neural-networs
[Richards G et al 2001]
Richards G, Rayward-Smith VJ, Sonksen PH, Carey S, Weng C. Data Mining for Indicators of Early Mortality in a Database of Clinical Records. Artif Intell Med 2001; 22: 215 – 231.
[Rokach L et al. 2008]
Rokach L, Maimon O. Data Mining with Decision Trees: Theory and Applications. Singapore: World scientific publishing co. pte.ltd; 2008.
[Rybchynski 2005]
Rybchynski D. Design of an Artificial Neural Network Research Framwork to Enhance the Development of Clinical Prediction Models. MASc Thesis. School of Information Technology and Engineering, University of Ottawa, Ottawa Ontario 2005.
[Samanta B et al. 2009]
Samanta B, Bird GL, Kuijpers M, et al. Prediction of Periventricular Leukomalacia. Part I. Selection of Hemodynamic Features Using Logistic Regression and Decision Tree Algorithms. Art Intel Med 2009;1058:1-15.
[Scheetz LJ et al. 2009]
Scheetz LJ, Zhang J, Kolassa J. Classification Tree Modeling to Identify Severe and Moderate Vehicular Injuries in Young and Middle-aged Adults. Artificial Intelligence in Medicine 2009;45: 1-10.
[Shang JS et al. 2000]
Shang JS, Lin YE, Goetz AM. Diagnosis of MRSA with Neural Networks and Logistic Regression Approach. Health Care Manag Sci 2000;3(4):287-97.
[Shi YP 2004]
Shi YP. Development of a Model for Prediction of Repeat Childhood Injuries in Injured Patients Using Artificial Neural Networks 2004;
[SMARTRISK. 2009]
SMARTRISK. The Economic Burden of Injury in Canada. SMARTRISK: 2009. Toronto, ON

[Statsoft Inc. 2008]
Statsoft Inc. Generalized Linear Models (GLZ). http://www.statsoft.com/TEXTBOOK/stglz.html.
[Swets JA 1988]
Swets JA. Measuring the Accuracy of Diagnostic Systems. Science. 1988; 240:1285–1293.
[Swets JA et al 2000]
Swets JA, Dawes RM, Monahan J. Better decision through science. Scientific American 283, 82-87.
[Taylor BJ 2006]
Brian J. Taylor. Methods and Procedures for the Verification and Validation of Artificial Neural Networks. Boston, MA, USA: Kluwer Academic Publishers, 2006. p 8.
[Tong Y et al. 2002]
Tong Y, Frize M, Walker R. Extending Ventilation Duration Estimations Approach From Adult to Neonatal Intensive Care Patients Using Artificial Neural Networks. IEEE Trans Info Technol Biomed 2002;6(2):188–91.
[Townsend D 2007]
Townsend D. Clinical Trial of Estimated Risk Stratification Prediction Tool. MASc Thesis. School of Information Technology and Engineering, University of Ottawa, Ottawa Ontario 2007.
[Tu JV et al. 1996]
Tu JV. Advantages and Disadvantages of Using Artificial Neural Networks Versus Logistic Regression for Predicting Medical Outcomes. J.Clin.Epidemiol. 1996;49(11):1225-31.
[Vergouwe Y et al. 2002]
Vergouwe Y, Steyerberg EW, Eijkemans MJ, Habbema JD. Validity of Prognostic Models: When is a Model Clinically Useful? Semin Urol Oncol 2002;20(2):96-107.
[Zhang G et al. 1998]
Zhang G, Berardi V. An Investigation of Neural Networks in Thyroid Function Diagnosis. Health Care Manag Sci 1998;1:29-37.
[Walker R et al. 1999]
Walker R, Frize M, Tong Y. Data Analysis Using Artificial Neural Networks and Computer-aided Decision-making in the Neonatal Intensive Care Unit. Paediatr Res 1999;45:231A.
[Wang MY et al. 2001]
Wang MY, Kim KA, Griffith PM, Summers S, McComb JG, Levy ML, Mahour GH. Injuries From Falls in the Pediatric Population: an Analysis of 729 Cases. J Pediatr Surg 2001 Oct; 36(10): 1528-1534.
[WHO on-line]
World Health Organization, Violence and Injury Prevention and Disability Department. Falls. http://www.who.int/violence_injury_prevention/other_injury/falls/en
[Wikipedia on-line]
http://upload.wikimedia.org/wikipedia/comments/e/e4/

Artificial_neural_networks.svg

[Williams S. et al]

William S. Davis and David C. Yen. The Information System Consultant's Handbook: Systems Analysis and Design. Chapter 28. CRC press 1998.

[Williamson LM et al. 2002]

Williamson LM, Morrison A, Stone DH. Trends in Head Injury Mortality Among 0-14 Years Olds in Scotland (1986-95). J Epidemiol Community Health. 2002,56:285-288.