STROKE RISK PREDICTION MODEL
BASED ON DEMOGRAPHIC AND MEDICAL SCREENING DATA

ACTING SUB LT. TEERAPAT KANSADUB

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR
THE DEGREE OF MASTER OF SCIENCE
(INFORMATION TECHNOLOGY MANAGEMENT)
FACULTY OF GRADUATE STUDIES
MAHIDOL UNIVERSITY
2016

COPYRIGHT OF MAHIDOL UNIVERSITY

Thesis
entitled

STROKE RISK PREDICTION MODEL BASED ON DEMOGRAPHIC AND MEDICAL SCREENING DATA

Teerapat Kansadub
.....
Acting Sub Lt. Teerapat Kansadub
Candidate

.....
Lect. Sotarat Thammaboosadee,
Ph.D. (Information Technology)
Major advisor

Chutima Jalayondeja
.....
Asst. Prof. Chutima Jalayondeja,
Dr.P.H. (Epidemiology)
Co-advisor

.....
Asst. Prof. Supaporn Kiattisin,
Ph.D. (Electrical and Computer
Engineering)
Co-advisor

.....
Prof. Patcharee Lertrit,
M.D., Ph.D. (Biochemistry)
Dean
Faculty of Graduate Studies
Mahidol University

.....
Asst. Prof. Supaporn Kiattisin,
Ph.D. (Electrical and Computer
Engineering)
Program Director
Master of Science Program in
Information Technology Management
Faculty of Engineering
Mahidol University

Thesis
entitled

STROKE RISK PREDICTION MODEL BASED ON DEMOGRAPHIC AND MEDICAL SCREENING DATA

was submitted to the Faculty of Graduate Studies, Mahidol University
for the degree of Master of Science (Information Technology Management)
on
March 30, 2016

Teerapat kansadub
.....
Acting Sub Lt. Teerapat Kansadub
Candidate

Wattana Jalayondeja
.....
Asst. Prof. Wattana Jalayondeja,
Ph.D. (Ergonomics/ Biomechanics)
Chair

Sotarat Th.
.....
Lect. Sotarat Thammaboosadee,
Ph.D. (Information Technology)
Member

.....
Asst. Prof. Supaporn Kiattisin,
Ph.D. (Electrical and Computer
Engineering)
Member

Chutima Jalayondeja
.....
Asst. Prof. Chutima Jalayondeja,
Dr.P.H. (Epidemiology)
Member

Kairoek Choeychuen
.....
Asst. Prof. Kairoek Choeychuen,
Ph.D. (Electrical and Computer
Engineering)
Member

.....
Prof. Patcharee Lertrit,
M.D., Ph.D. (Biochemistry)
Dean
Faculty of Graduate Studies
Mahidol University

.....
Asst. Prof. Jackrit Suthakorn,
Ph.D. (Robotics)
Dean
Faculty of Engineering
Mahidol University

ACKNOWLEDGEMENTS

This thesis would not be success without advices, supports and encouragements from many people. I have to thank my research supervisors, Dr. Sotarat Thammaboosadee, Asst. Dr. Prof. Supaporn Kiattisin, and Asst. Prof. Dr. Chutima Jalayondeja, who suggest me on research of this topic and knowledge of data mining, and thinking process. I would like to thank you very much for your support and understanding over these past two years.

Most importantly, none of this could have happened without my family, my friends, and colleague who always stays beside and encourage me when I was under pressure and discouraged. Although they do not know deeply in this topic, they tend to support me as much as possible. Finally, I have to thank Asst. Prof. Dr. Prasert Sakulsriprasert who advices about English articles writing.

Acting Sub Lt. Teerapat Kansadub

STROKE RISK PREDICTION MODEL BASED ON DEMOGRAPHIC AND MEDICAL SCREENING DATA

ACTING SUB LT. TEERAPAT KANSADUB 5736274 EGIT/M

M.Sc. (INFORMATION TECHNOLOGY MANAGEMENT)

THESIS ADVISORY COMMITTEE: SOTARAT THAMMABOOSADEE, Ph.D., CHUTIMA JALAYONDEJA, Dr. P.H., SUPAPORN KIATTISIN, Ph.D.

ABSTRACT

Nowadays, strokes are the third leading cause of Thai's mortality in all age groups. The statistical data during 1994 - 2013 found that strokes caused 255,307 mortalities. In this paper, we present the data mining model for stroke prediction to screen people having strokes. Three classification algorithms including Neural Network, Decision Tree, and Naïve Bayes, are used for stroke prediction with different datasets: demographic data, medical screening data, and integrated data. This research was initialized with attributes and data selection, data collection, data resampling, data integration, data grouping, modeling, evaluation, and deployment. The best experimental result is Neural Network applied with integrated data result with 0.84 accuracy, 0.12 false positive rate, 0.25 false negative rate, and 0.9 area under ROC curve (AUC). Furthermore, the factor analysis using the best integrated data based on decision tree found that hemophilia and balance loss are the new, discovered risk factors compared with prior research. Finally, the best model was also used to develop an application for user-friendliness.

KEY WORDS: STROKE/ DATA MINING/ PREDICTION/
DEMOGRAPHIC DATA

60 pages

การทำนายปัจจัยเสี่ยงโรคหลอดเลือดสมองโดยใช้ข้อมูลพื้นฐาน และ ข้อมูลแบบคัดกรองคนไข้
STROKE RISK PREDICTION MODEL BASED ON DEMOGRAPHIC AND MEDICAL SCREENING DATA

ว่าที่ ร.ต. ธีรพัฒน์ กันสดับ 5736274 EGIT/M

วท.ม. (การจัดการเทคโนโลยีสารสนเทศ)

คณะกรรมการที่ปรึกษาวิทยานิพนธ์ : โษฑศ์รัตต ธรรมบุษดี, Ph.D., ชุติมา ชลายนเดชะ, Dr. P.H., สุภาภรณ์ เกียรติสิน, Ph.D.

บทคัดย่อ

ทุกวันนี้โรคหลอดเลือดสมอง เป็นสาเหตุที่ทำให้คนไทยเสียชีวิตเป็นอันดับสาม ในทุกช่วงอายุ จากสถิติระหว่างปี 2537 -2556 พบผู้เสียชีวิตจากโรคหลอดเลือดสมองจำนวน 255,307 ราย ในวิทยานิพนธ์นี้จึงได้นำวิธีการวิเคราะห์ข้อมูล เช่นการทำเหมืองข้อมูลเข้ามาสร้างแบบจำลองในการทำนายการเกิดโรคหลอดเลือดสมอง โดยในงานวิจัยนี้ใช้สามอัลกอริธึมในการสร้างแบบจำลอง ได้แก่ โครงข่ายประสาทเทียม ต้นไม้ตัดสินใจ และ เบย์อย่างง่าย เพื่อใช้ในการทำนายการป่วยเป็นโรคหลอดเลือดสมอง โดยวิธีการเหล่านี้จะถูกนำไปใช้กับชุดข้อมูลที่แตกต่างกันได้แก่ ข้อมูลพื้นฐานคนไข้ ข้อมูลแบบคัดกรองคนไข้ และข้อมูลแบบบูรณาการ โดยการศึกษาครั้งนี้เริ่มจากการเลือกปัจจัยที่ใช้ในการทำนาย การเก็บข้อมูลพื้นฐาน และ ข้อมูลแบบคัดกรอง การสุ่มกลุ่มตัวอย่าง การบูรณาการข้อมูล การจัดกลุ่มของข้อมูล การสร้างแบบจำลอง การประเมินผล และ การนำไปใช้งาน โดยผลที่ได้รับคือวิธีโครงข่ายประสาทเทียมเมื่อใช้ร่วมกับข้อมูลแบบบูรณาการเป็นแบบจำลองที่ได้ผลดีที่สุดโดยมีค่าความแม่นยำเท่ากับ 0.84 False Positive เท่ากับ 0.12 False Negative เท่ากับ 0.25 และ พื้นที่ใต้โค้งROC (AUC) เท่ากับ 0.9 นอกจากนี้เมื่อนำปัจจัยจากวิธีต้นไม้ตัดสินใจที่ดีที่สุดมาวิเคราะห์ พบว่าปัจจัยโรคฮีโมฟีเลีย และ สูญเสียการทรงตัว เป็นปัจจัยเสี่ยงที่ถูกค้นพบใหม่เมื่อเปรียบเทียบกับงานวิจัยอื่นก่อนหน้านี้ ในขั้นตอนสุดท้ายแบบจำลองจากโครงข่ายประสาทเทียมร่วมกับข้อมูลแบบบูรณาการได้ถูกนำมาประยุกต์ในการสร้างแอพพลิเคชันเพื่อความสะดวกในการใช้งาน

60 หน้า

CONTENTS

CONTENTS (cont.)

CONTENTS (cont.)

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
ANN	Artificial Neural Network
AUC	Area under the Curve
BMI	Body Mass Index
CAD	Coronary Artery Disease
CRISP	Cross Industry Standard Process
CVD	Cardiovascular Disease
ER Diagram	Entity-Relationship Diagram
FN	False Negative
FP	False Positive
HN	Hospital Number
ICD	International Classification of Diseases
IRB	Institutional Review Board
MLP	Multilayer Perceptron
NESDB	National Economic and Social Development Board
ROC	Receiver Operating Characteristic
SD	Standard Deviation
WHO	World Health Organization

CHAPTER I
INTRODUCTION

Nowadays, stroke disease is a serious public health problem and is the third largest cause of death and needed for identification or diagnostic system to be essential for pre-detected patients. Therefore, this thesis purposes methodology for creating stroke risk prediction model and its deployment.

1.1 Background and Statement of Problems

Stroke is caused by abnormal of vascular in brain and affecting of neuro such as muscle weakness, numb, and probably mortality harm. Stroke can be separated into two types: Ischemic stroke and Hemorrhagic strokes (Poungvarin, 2001). Ischemic stroke is the most common cause of disease. It is a result of occlusion of the major cerebral arteries. Hemorrhagic strokes occur as a result of leaky to blood vessels with bleeding into the brain tissue. Generally, stroke will impact in your daily-life such as memories, movement, visuality, speaking and read-write ability.

Statistic of World Health Organization (WHO) stated that stroke is the third leading cause of mortality in females and males of all life periods (WHO, 2015). Interestingly, statistic of Office of the National Economic and Social Development Board (NESDB) during 1994 to 2013 also reported that stroke causes mortality 255,307 cases (NESDB, 2015).

To define symptom of stroke is tedious and time-consuming for medical diagnosis. Therefore, it needs for the automate system to predict risk of stroke which have various types of data such as Demographic, Medical Screening data, etc. Currently, there are several existing prediction methods of stroke disease but those methods still have limitations. They use enormous factors and data source such as medical history and the symptoms. However, researcher supposes that the Demographic, Medical

Screening, and their integration are sufficient to create prediction model driven by data mining technique.

1.2 Objective of Study

The objective of this thesis is to develop the stroke risk prediction model based on two patients' information including demographic data and medical screening data. The smashing model is selected from three classification data mining algorithms, given as Artificial Neural Network, Naïve Bayes, and Decision Tree. In addition, the factors analysis results are also explored and compared with the previous research. An application is also extensively developed for ease of use for naïve users.

1.3 Scope of Work

This thesis is to develop a new predict method for classification and prediction by Researcher who collected the stroke patients, demographic and medical screening data, from Faculty of Physical Therapy during from 2012 to 2015. The selection criteria is stated in Chapter 3.

1.4 Expected benefits

The expected benefits of this research are the stroke risk prediction model based on demographic and medical screening data. This model will be assured with the technical and medical aspects.

CHAPTER II
RELATED THEORIES AND RESEARCHES

This chapter consists of theories related to impact of stroke, stroke risk factors, inclusion criteria, exclusion criteria, international classification of diseases for classification, and concept of data mining. Moreover previous related works are also reviewed.

2.1 Stroke

2.1.1 Impact of stroke

Stroke is a major public health care concern and has a significant impact on individuals, families and wider society. In UK, there are the estimated 150,000 people getting stroke each year (Office of National Statistics, 2001). Stroke is the third most common cause of death, after the heart disease and cancer, with over 67,000 deaths each year (Townsend, 2005). However, the most significant and lasting impact of stroke is long-term disability. Stroke is the greatest cause of complex and severe adult disability in the UK (Wolfe, 2000) (Adamson *et al.*, 2004). A third of stroke patients have some long-term disability (National Audit Office (NAO), 2005). Common consequent problems of stroke include aphasia, physical disability, loss of cognition and communication skills, depression, and other mental health problems.

Stroke is significant on health services. In England, it costs the National Health Service (NHS) to £7 billion per year (NAO, 2005). Unfortunately, outcomes in the UK compare poorly internationally although services most expensive and disability or mortality (Leal *et al.*, 2006).

2.1.2 The important stroke risk factors

According to Epidemiological data shown in international INTERSTROKE study (Poungvarin, 2001), there are nine modifiable factors found. These latter factors are reviewed as follows.

1) Hypertension

Hypertension is high blood pressure symptom by measuring over 140/90 mmHg. This risk factor could be considered as heredity, diabetes, and obesity. It usually increases risk of stroke by four to six times. Hypertension leads to atherosclerosis and hardening of the large arteries. This can lead to blockage of small blood vessels in brain. High blood pressure brings to broken or leak in brain. The risk of stroke is directly related to the high blood pressure happen.

2) Diabetes

Diabetes is caused by high sugar level blood over 126 mg/dL. It begins with the pancreas does not contain enough insulin or cells do not response with insulin. Diabetes is secondary risk factor from hypertension and related with hypertension that is risk factor of stroke.

3) Obesity

Obesity is condition of body fat when the body mass index (BMI) calculated by the division of height and weight is over 30 kg/m^2 defined as overweight.

4) Dyslipidemia

Dyslipidemia (Longo *et al.*, 2008) is abnormal amount of lipids in the blood. Diagnosis is given by measuring the plasma levels of total cholesterol, TGs, and individual lipoproteins. Dyslipidemia makes blood vessel be constricted effect with blood flow. Lacking of blood to support internal organs may paralyze the brain.

5) Smoking

Smoking may lead to stroke by increasing platelet agreeability, blood viscosity, fibrinogen levels, or by causing vascular damage.

6) Cardiovascular disease

Cardiovascular disease (CVD) is a category of disease including heart and blood vessels. CVD includes coronary artery disease (CAD) likes angina and myocardial infarction. Other CVDs are stroke, heart arrhythmia, vascular heart disease,

peripheral artery disease, cardiomyopathy, venous thrombosis, aortic aneurysms, peripheral artery disease, and rheumatic heart disease.

7) Medicine

Some drugs are a common stroke risk factor. It induces stroke by damaging of the blood vessels. Those drugs are such as amphetamines and cocaine.

8) Age

Age is important risk factor of cardiovascular or heart disease. It is estimated to 82 percent who mortality of coronary heart disease by blood vessel cannot support blood pressure so maybe leak or broken.

2.2 Inclusion Criteria

Inclusion criteria for choosing population is identified by diagnosis code (ICD10) between I60 (Nontraumatic subarachnoid hemorrhage) and I69 (Sequelae of cerebrovascular disease) (Poungvarin, 2001).

2.3 Exclusion Criteria

Exclusion criteria are patients who are older than 20 years old for significant analysis (Poungvarin, 2001).

2.4 International Classification of Diseases (ICD)

The International Classification of Diseases (ICD) is the standard tool for diagnosis in epidemiology, health management and clinical objective. This includes the analysis of the general health, incident monitor, and prevalence of disease. ICD is used by physicians, nurses, providers, and researchers to classify diseases and other health problems recorded by ICD codes Finally, ICD is used for reimbursement and resource allocation decision-making also in some country.

The ICD standard has been continuously developed and published in several versions. In this thesis, the ICD-10 standard is used as a standard data collection format. ICD 10 is code character mix with number (Alphanumeric code) by each character starting from at A to Z (WHO, 1994). Approaching to the stroke diagnosis, the code range is scoped to I60-I69.

2.5 Data mining

2.5.1 Data mining definition

To understand the term of data mining, data mining is the knowledge extracting from the mass of data that are unnoticed. From aspect of scientific research, data mining is related to the new rules of study, including computing, marketing, and statistics. Many algorithms used in data mining come from two research communities: machine learning community and statistical community. Particularly in multivariate and computational statistics, machine learning is related with computer science and artificial intelligence (AI) and is about finding relations and regularities from data. The aim of machine learning is analysts and extract unobserved cases.

The data mining is to obtain the results that can be measured in terms of their relevance for the owner of database as a business advantage. Here is a more complete definition of data mining (Giudici, 2003):

"Data mining is the process of selection, exploration, and modelling of large quantities of data to discover regularities or relations that are at first unknown with the aim of obtaining clear and useful results for the owner of the database."

In business, the data mining is to analyze the data in database aims to extract knowledge from database to the business needs by applying a machine learning or statistical technique implemented in a computer algorithm. and in the final achieve by receive strategy decision. The strategic decision will itself create new measurement needs and consequently new business needs, setting off what has been called "the virtuous circle of knowledge" induced by data mining (Berry and Linoff, 1997).

2.5.2 The data mining process

The data mining process have six stages, called Cross Industry Standard Process (CRISP) (Chapman *et al.*, 2000), as shown in Figure 2.1.

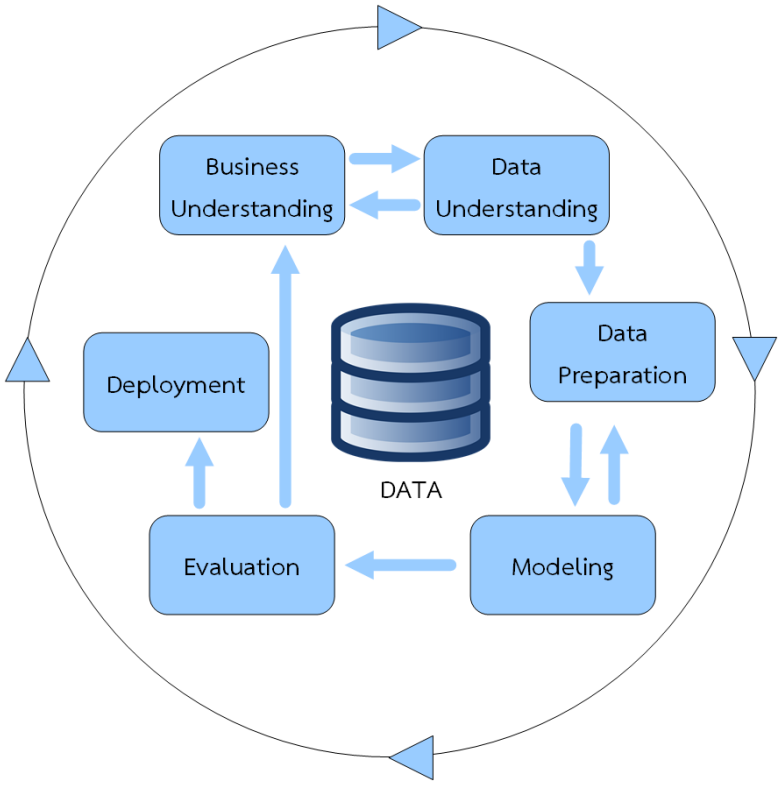


Figure 2.1 The CRISP-DM life cycle.

- The first step is business understanding. This phase is to understand about business problem and opportunity.
- The second step is data understanding. This phase relates with data collection. Under data gathering the data should appropriate with data mining method.
- The third step is data preparation. This step is very important because the quality of model depend on data quality by related process such as cleaning data, replace data, grouping data and data arranged.
- The fourth step is data modeling. This step is to creat model by using training set of data for create model. Algorithms are such as Neural Network, Decision Tree and Naïve Bayes.

- The fifth step is data evaluation. This phase relates with evaluation of model performance by one of the popular method for evaluate is 10-fold cross validation.
- The final step is deployment. This method has various style. One of them is to develop application based on discovered model and to use it to evaluate new data and receive result from prediction in application.

2.5.3 Data Sampling

In this research, data sampling is used to solve the imbalance problems. Under-sampling is to randomly remove majority class. Over-sampling is to help achieving balance class distribution by replication minority class sample or combining it together (Longadge, Dongre, and Malik, 2013).

2.5.4 Normalization

When approaching data modeling some standard should apply to prepare data to data modeling as normalize to bring all variables into one to adjust for the disparity in the variable sizes by Normalize value can calculate from and result between 0 and 1 (Roiger, 2003). Benefit of normalization value for compare how difference between original dataset and dataset after resampling in order to find which resampled dataset similar with original dataset.

2.5.5 Artificial Neural Network

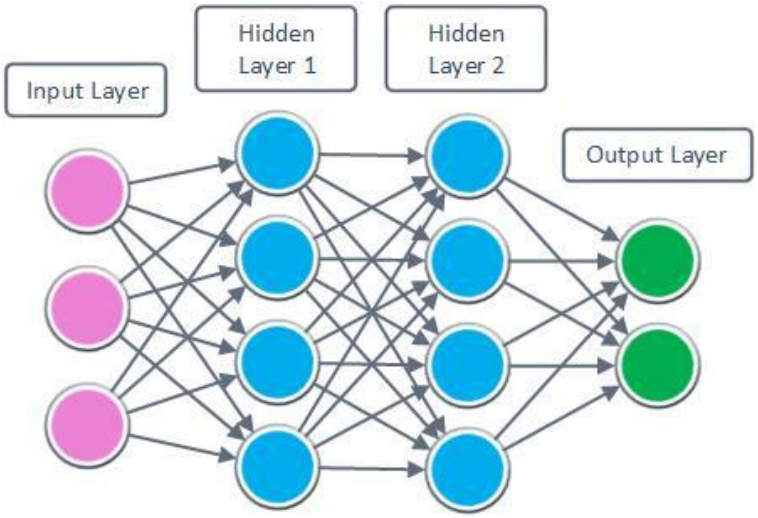


Figure 2.2 Example of Neural network method.

Artificial Neural Network (ANN) (Duda, Hart and Stork, 1973) is simulation of the brain function of human by consist of input layer, hidden layer, and output layer. It also consists of several adjustable parameters.

Hidden layer is a black box in order to transform inputs into something and go to output layer by amount hidden layer can adjust various styles. Learning rate and momentum are the widely tunable parameters in the backpropagation algorithm. When the gradient of the error function keep pointing in the same direction this increase size of step, it is necessary to reduce the global learning rate when high value of momentum. Over value of learning rate and momentum may cause the tuning process pass the minimum error point while the model can be stopped at the local minima if they are too small. Output layer is class that check prediction answer true or false. Activation function is calculation function in hidden layer receive signal from previous layer by each layer may be difference function.

2.5.6 Decision Tree

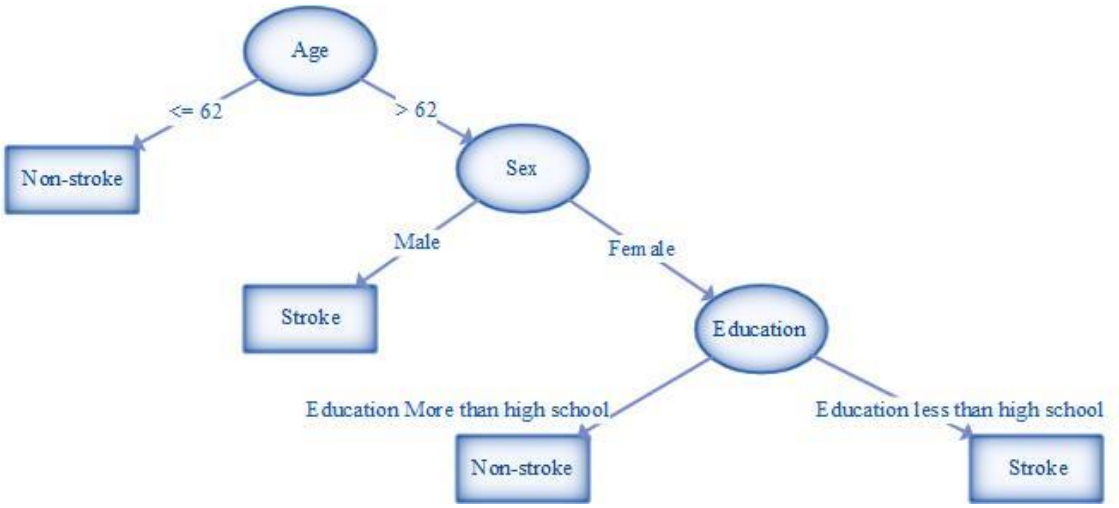


Figure 2.3 Example of Decision tree method.

The decision tree (MacQueen, 1967) is a tree similar flow chart which consists of an element as root node, internal nodes, leaf nodes, and branches. The root node is represented as the top of chart that is major factor that can the best cluster data, initiate with create the tree by training data and extract a node to leaf node then recombines and removes the branch that affects the accuracy. C4.5 (Quinlan, 1993) is an algorithm generate a decision tree develop by Ross Quinlan by C4.5 is an algorithm for classifier and build decision tree from training data set each node of tree, C4.5 chooses attribute highest normalized information gain to make the decision.

2.5.7 Naïve Bayes

The Naïve Bayes classifier offers the supervised classification with all input attribute equal importance and independent of one another as shown in the Equation 1.

$$P(H|E) = \frac{P(E|H) \times P(H)}{P(E)} \qquad (1)$$

Where H is hypothesis to be test and E is evidence associated with hypothesis by hypothesis independent variable and represented class, and evidence is determined by the values of input attributes. P(H|E) is the condition probability by H is true given evidence E. P(H) is priori probability represent probability of hypothesis before representation any evidence (Roiger, 2003).

2.5.8 Model Evaluation

1) 10-Fold Cross Validation

Ten Fold Cross Validation (Kohavi, 1995) is a technic aims to measure the result of the prediction by split a dataset into ten and nine proportion are used as training set and the rest is used for testing. The training will be performed in 10 times as shown in Figure 2.4.

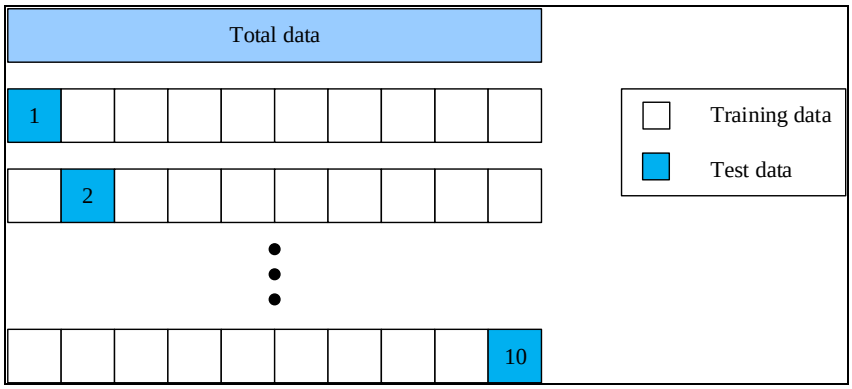


Figure 2.4 Ten Fold Cross Validation technic.

2) Accuracy, False Positive rate, and False Negative rate

Accuracy (Zhu and Davidson, 2007) is one of the method for measures the correct prediction ratio by use True Negative value, True Positive value, False Negative value and False Positive value for calculation.

False Positive rate (FP rate) (Kohavi and Provost, 2001) is value proportion in negative section predict cases were incorrectly classified as positive by FP rate calculated from false positive and true negative value.

False Negative rate (FN rate) (Kohavi and Provost, 2001) is value proportion in positive section predict cases were incorrectly classified as negative by FN rate calculated from false negative and true positive value.

3) Area Under Curve (AUC)

In prediction have graph receiver operating characteristic (ROC) curve that measures the predictive accuracy of model based on confusion matrix. AUC is area under the ROC curve so AUC maybe effect with accuracy (Giudici, 2003).

2.6 Related works

1) Effective Analysis and Predictive Model of Stroke Disease using Classification Methods (Sudha, Gayathri & Jaisankar, 2012)

This research proposes predictive model of stroke. It consists of attributes containing medical history and symptoms of patients such as physical exam results, blood test results, and diagnoses. In the first step, dataset of stroke is collected from medical therefore data preprocessing by removes duplicate records, missing data, noisy and inconsistent data. Dimensional Reduction is processed by PCA algorithm and Clustering. This work studies by comparing accuracy, False Positive (FP), False Negative (FN), and Area Under Curve (AUC) in order to convert to rule and result from three methods between Decision tree, Naïve Bayes, and Neural Network and explain the best rule for benefits of population who have risk in stroke. The result showed that the decision tree is the best classification algorithm for studying of predicting cerebrovascular diseases by achieves 95.29% of sensitivity and 98.01% of accuracy.

Although in the result achieves high value of accuracy, in prediction do not category of data.

2) Prevalence of Stroke and Stroke Risk Factor in Thailand (Hanchaiphiboolkul *et al.*, 2011)

In this analysis, the authors did cross-sectional analysis by using baseline health survey data. Age, gender, and region-specific prevalence for stroke were derived, and 95% confidence intervals were calculated. The crude rates were standardized to Segi world standard population (18) and the new WHO world standard population (18). Regarding to demographic data and risk factors, continuous variables were presented as mean and standard deviation (SD). Categorical variables were described with percentages. A univariate logistic model was used to examine the individual relationship between each variable and stroke. After each variable was tested independently in a univariate regression model, those achieving a p-value < 0.20 were selected for testing in multivariate logistic regression. Odds ratio (ORs) and 95% confidence intervals (CIs) were used to illustrate the association between potential risk factors and stroke.

This research used multiple logistic regression method analysis factors associated with stroke prevalence. Although this thesis use demographic and

data associate with stroke risk factor for analyze, the final result this method does not create model and cannot be deployed to application likes the data mining method.

3) Sex Differences in First-Ever Acute Stroke (Roquer, Campello and Gomis, 2003)

Sex determines some clear differences in patients suffering a first-ever stroke. Women were, on average, 6 years older than men and had a different profile of vascular risk factors and a different distribution of stroke subtypes. Women had a longer hospital stay and remained more disable than men. The amelioration of hypertension control and increase in anticoagulant treatment in patients with atrial fibrillation would be the best options for preventing stroke, especially in women. Although statistical was used in this research, there were many risk factors compared in couple so some factor maybe do not compare thorough as data mining methods.

According to reviewed theories including researches, researcher separates the experiment into three dataset: demographic dataset, medical screening dataset and integrated dataset with three methods, including Neural Network, Decision Tree and Naïve Bayes, as describe in next chapter

CHAPTER III
RESEARCH METHODOLOGY

The research methodology of this research is shown in Figure 3.1.

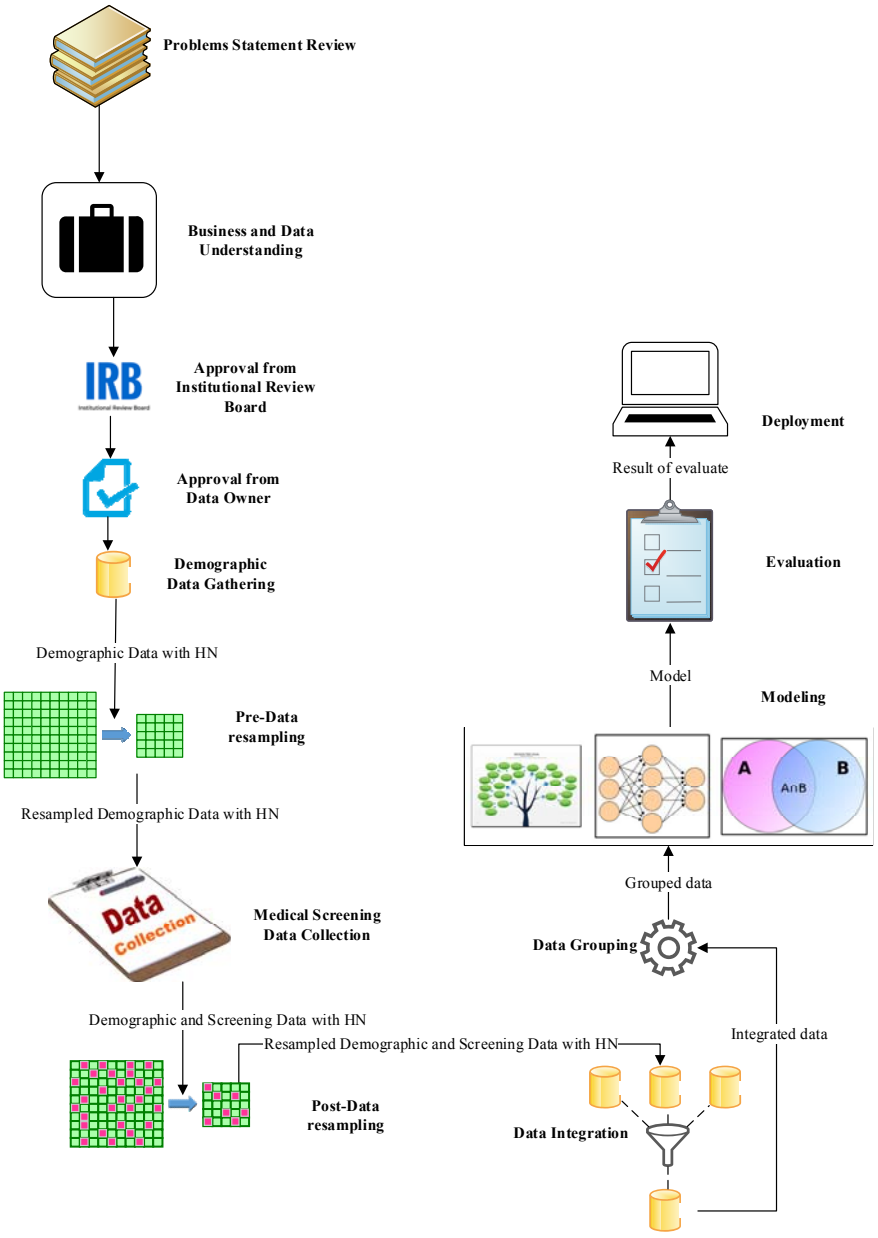


Figure 3.1 Flow chart of research methodology.

To describe research methodology in detail, each research step is stated in each section.

3.1 Problems Statement Review

As described in Chapters 1 and 2, nowadays stroke is one of the top disease mortality of patients align from statistic of world health organization (WHO) (WHO, 2015). Therefore, researcher awares that it is important to find the risks factors and their identification procedure to prevent the severity earlier.

3.2 Business and Data Understanding

According to business and data understanding, researchers has studied the business process in the services of Faculty of Physical Therapy services about patient treatment. Initially, the study found that patient having orthopedic and neuro disease are the stroke patients. The patients-related data are in two sources. The first one is demographic data which are the general information of patients electronically stored in database. Another one is medical screening data which are recorded in paper-based format.

3.3 Approval from Institutional Review Board

Before gathering the data, the job of research must request for approval from Institutional Review Board (IRB) in order to legally align the ethic of research. The approval document is attached in Appendix B.

3.4 Approval from Data Owner

Since IRB has approved the research, researcher requested for approval from data owner, the Faculty of Physical Therapy, as shown in document appeared in Appendix B.

3.5 Demographic Data Gathering

After approval from the IRB and Data Owner, the demographic data are gathered from database by SQL querying. The source of data is from medical institute in the Faculty of Physical Therapy, Mahidol University during 2012 - 2015. The gathered factors are sex, age, province, marital status, education, and occupation. The stroke and non-stroke patients data are discriminated by ICD-10 code in a range of I60-I69. The patients information consists of several relationships and attributes. The data dictionary of all tables are shown in Tables 3.1 - 3.7, and the Entity-Relationship diagram (ER diagram) of all tables are shown in Figures 3.2 - 3.4. According to exclusion criteria stated in Chapter 2, the data gathering process is executed by filtering out the patients who are less than 20 years old (Poungvarin, 2001).

Table 3.1 Data dictionary of Province Table.

No.	Attribute	Description
1.	Province_ID	ID No. of province
2.	Province_Th	Name of province in Thai
3.	Province_Eng	Name of province in English
4.	IsEnable	Status of row

Table 3.2 Data dictionary of Address Table.

No.	Attribute	Description
1.	Address_ID	ID No. of address
2.	PT_ID	ID No. of patients
3.	AddMooName	Name of village
4.	AddNo	Home number
5.	AddMoo	Moo
6.	AddRoad	Name of road
7.	AddTumbon	Name of subdistrict
8.	AddDistrict	Name of district
9.	AddProvince_ID	ID No. of province
10.	AddZipcode	Zip code
11.	AddPhone	Phone number
12.	AddMobile	Mobile phone number
13.	EMTitle	Prefix of relative patient
14.	EMFirstName	First name of relative patient
15.	EMLastName	Last name of relative patient
16.	EmRelation	How relative with patient
17.	ChkAddType	Type of address each instance between Current address, Work address, and Emergency address
18.	Remark	Remark of address

Table 3.3 Data dictionary of Patient Table.

No.	Attribute	Description
1.	PT_ID	ID No. of patient
2.	Photo	Photo of patient
3.	Prefix	Prefix of patient
4.	Prefix_Rang	Rank in Thai
5.	FirstName	First name of patient in Thai

Table 3.3 Data dictionary of Patient Table (Cont.).

No.	Attribute	Description
6.	LastName	Last name of patient in Thai
7.	PrefixEng	Rank in English
8.	Prefix_RangEng	Prefix of patient in English
9.	FirstNameEng	First name of patient in English
10.	LastNameEng	Last name of patient in English
11.	Nickname	nickname of patient
12.	ID Card	ID card number
13.	Birthday	Birthday of patient
14.	AgeYear	Age of patient : Year
15.	AgeMonth	Age of patient : Month
16.	AgeDay	Age of patient : Day
17.	Sex	Gender of patient
18.	Nationlity_ID	ID No. of Nationality
19.	MaritalStatus	Marital Status of Patient
20.	MS_Other	Marital Status of Patient in Addition from Marital Status
21.	Stature	Height of Patient
22.	Weight	Weight of Patient
23.	PresentAdd_ID	ID No. of Address
24.	OfficeAdd_ID	ID No. of Office Address
25.	Emergency_ID	ID No. of Contact Person when emergency
26.	Occupation_ID	ID No. of occupation
27.	Occupation_Other	Other occupation
28.	Education_ID	ID No. of education
29.	Education_Other	Other education
30.	RightTreatment	Right Treatment
31.	RT_Other	Other Right Treatment
32.	RT_Institution	Institution Right Treatment

Table 3.3 Data dictionary of Patient Table (Cont.).

33.	ChkTreatment	Method Coming Treatment
34.	TreatmentName	Name Introduce
35.	TreatmentPlace	Place Introduce
36.	IsDelete	Status of row
37.	CreateDate	Date create row
38.	CreateBy	ID No. of member create row
39.	UpdateDate	Date Last Update row
40.	UpdateBy	ID No. of member update row
41.	RegisterDate	Date register
42.	Status_ID	Status of patient
43.	Branch_ID	ID No. of branch
44.	Remark	Remark of patient
45.	TypeOfAdd	Type of address

Table 3.4 Data dictionary of Medinfo_Diagnosis Table.

No.	**Attribute**	**Description**
1.	MedinfoDiag_ID	ID No. of medical info diagnosis
2.	PT_ID	ID No. of patient
3.	Medinfo_ID	ID No. of patient medical info
4.	Diagnosis_Code	Code of diagnosis
5.	TherapyType_ID	ID No. of therapy type
6.	Remarks	Remark of medical info diagnosis
7.	DiagnosisType_ID	ID No. of Diagnosis type

Table 3.5 Data dictionary of Nationality Table.

No.	Attribute	Description
1.	Nationality_ID	ID No. of nationality
2.	Nationality_Th	Name of nationality in Thai
3.	Nationality_Eng	Name of nationality in English
4.	IsEnable	Status of row

Table 3.6 Data dictionary of Occupation Table.

No.	Attribute	Description
1.	Occupation_ID	ID No. of occupation
2.	Occupation_Th	Name of occupation in Thai
3.	Occupation_Eng	Name of occupation in English
4.	IsEnable	Status of row

Table 3.7 Data dictionary of Education Table.

No.	Attribute	Description
1.	Education _ID	ID No. of education
2.	Education _Th	Name of education in Thai
3.	Education_Eng	Name of education in English
4.	IsEnable	Status of row

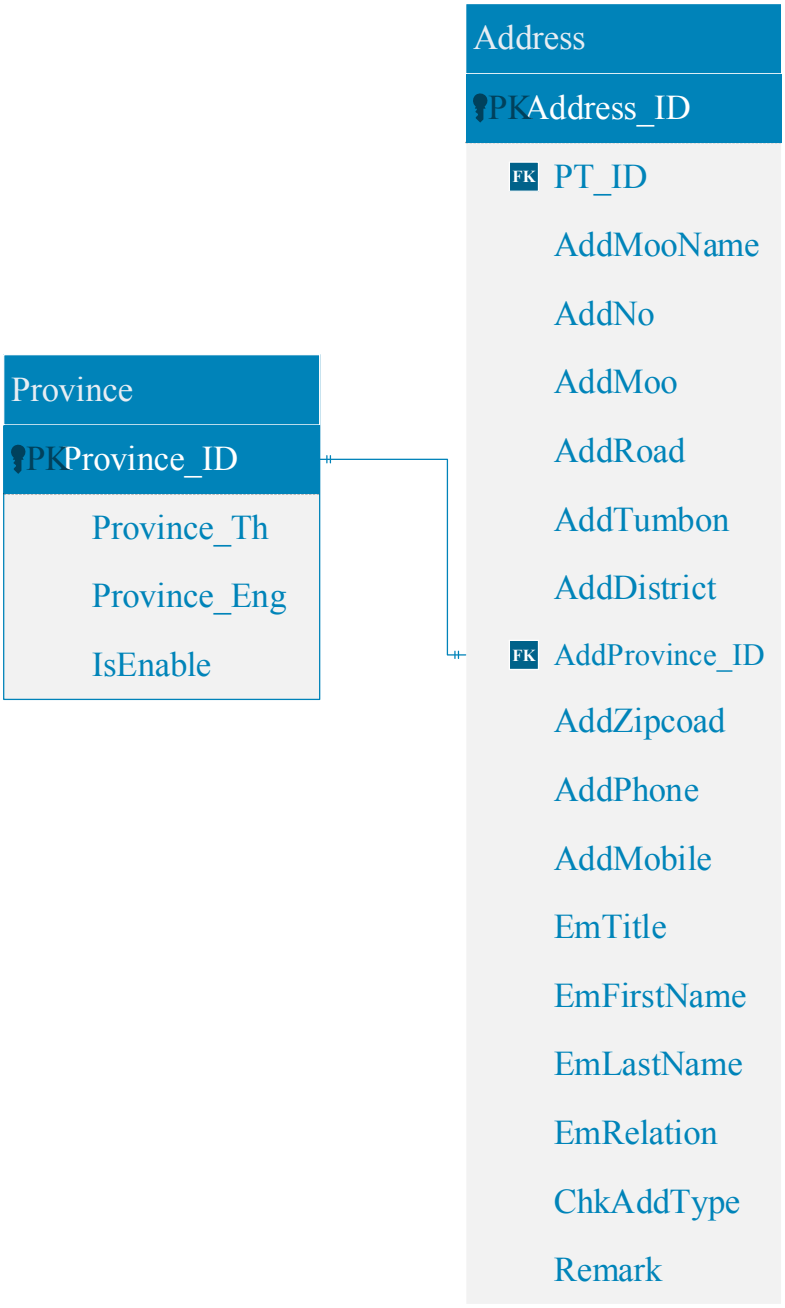


Figure 3.2 Entity- Relationship Diagram between Province and Address tables.

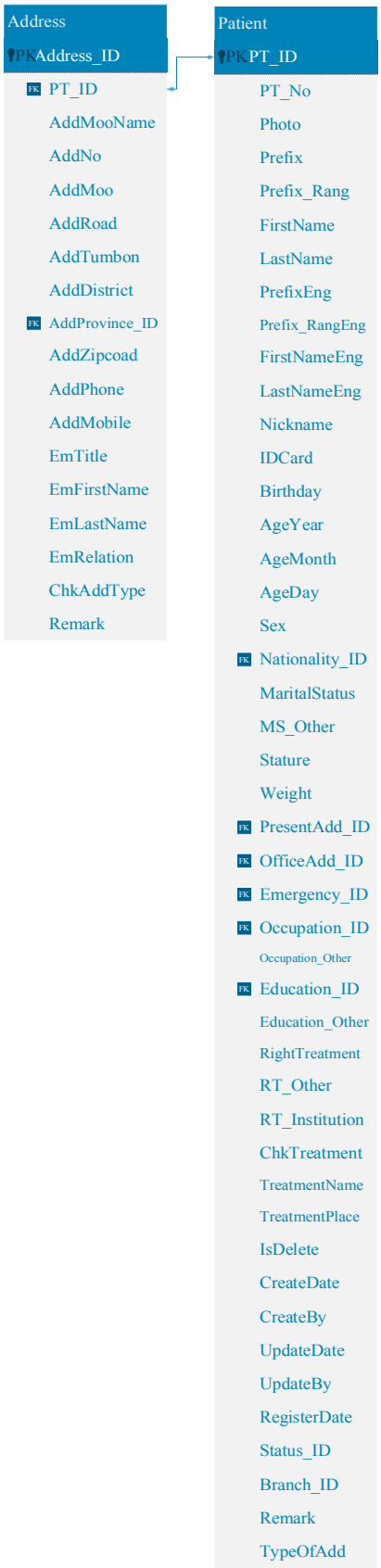


Figure 3.3 Entity- Relationship Diagram between Address and Patient tables.

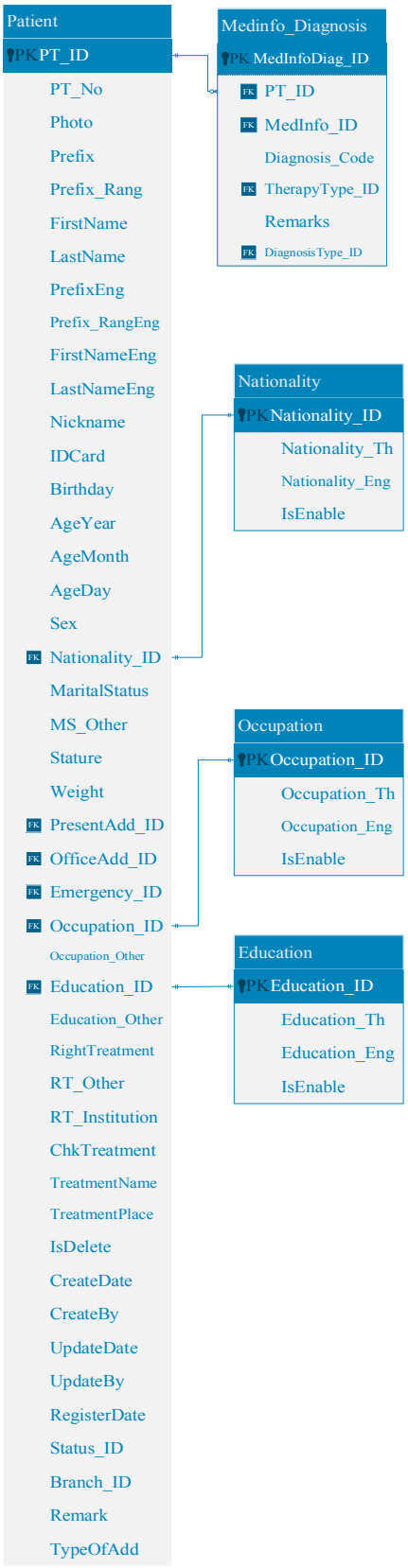


Figure 3.4 Entity- Relationship Diagram between patient's demographic and Medical diagnosis code tables.

3.6 Pre-Data Resampling

After data gathering, the imbalance of data amount was found (Longadge, Dongre, and Malik, 2013) (Prati, Batista, and Monard). Stroke and non-stroke patients data amount are in the ratio of 250:67,897. To balance the data proportion for the performance of discrimination, the researcher performed down-sampling for the non-stroke data to be reduced to 500 in order to maintain the ratio 1:2. The resampling has been processed for several seeds and use normalized euclidience distance to select the closet one compared with original dataset.

3.7 Medical Screening Data Collection

The medical screening data are required for advancement of data analysis upon the demographic information. Unfortunately, the stored patients data are only the demographic data while the medical screening data are collected in paper-based medical screening form. The selected patient records including H/N no., are used to manually retrieve from archival storage shelving located at the Medical Institute of Faculty of Physical Therapy, Mahidol University. To facilitate the gathering process, a simple macro-based spreadsheet application was developed for the data collection as illustrated in Figure 3.5.

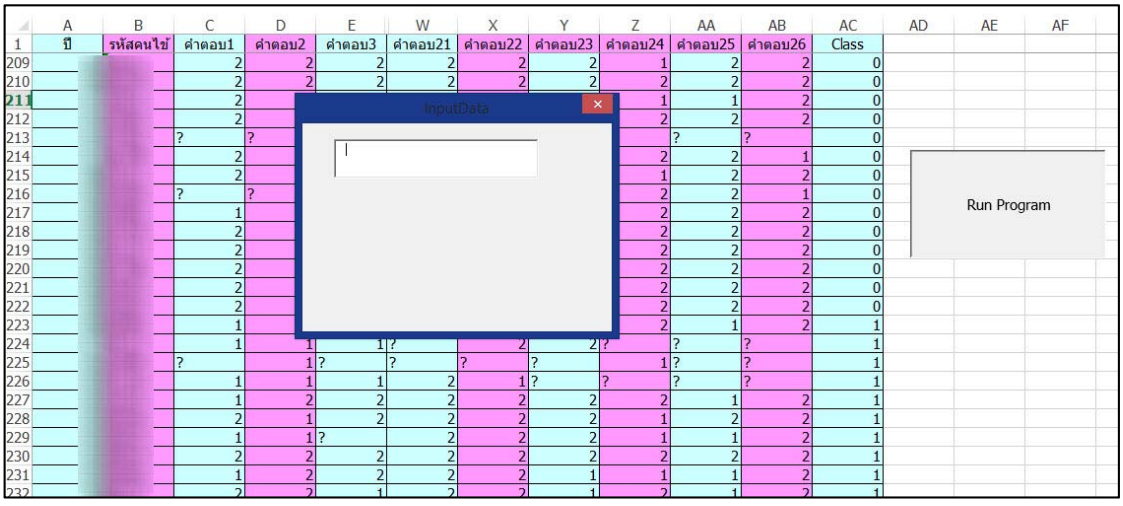


Figure 3.5 Tool for Medical Screening Data Collection.

3.8 Post-Data Resampling

After data collection was executed, the researcher found that some medical files of stroke patients are loss due to the loose management policy. Thus, to maintain the balance of data, the non-stroke patient records are down-sampled again which aims to the proportion of 2:1. Consequently, the final ratio between stroke and non-stroke patients are 294:147 by using normalized value procedure as applied in the previous data resampling phase.

3.9 Data Integration

Due to the several versions of medical screening forms developed by years, the collected data from four versions are merged to be single one, as illustrated in Figure 3.6.

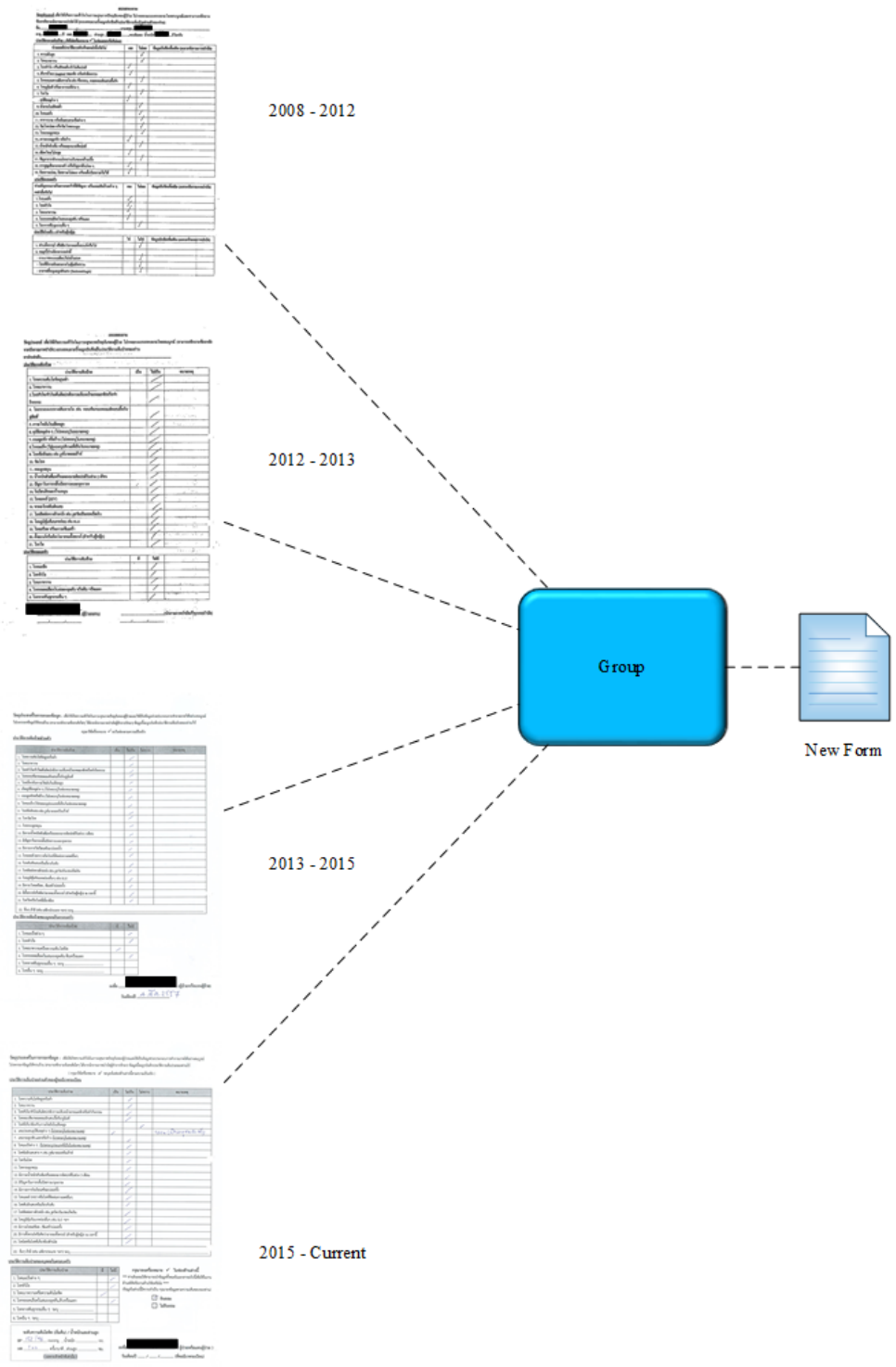


Figure 3.6 Medical Screening Form integration.

After medical screening form integration was gathered, researcher had found new medical screening data factors as shown in Table 3.8. The integrated form

consists of several factors with Boolean value. In this research, the encoding is given as 1 for true, -1 for false, and 0 for unknown or unspecified.

Table 3.8 Attributes of Medical Screening Data.

No.	Attribute	Description
1	Hypertension	The Boolean value indicates for high blood pressure.
2	Diabetes	The Boolean value indicates for diabetes.
3	Heart disease	The Boolean value indicates for heart Disease.
4	Asthma Bronchitis Allergy	The Boolean value indicates for asthma.
5	Hyperlipidemia	The Boolean value indicates for hyperlipidemia disease.
6	Accident	The Boolean value indicates for ever Accident.
7	Fracture	The Boolean value indicates for ever Fracture.
8	Cancer	The Boolean value indicates for cancer disease.

Table 3.8 Attributes of Medical Screening Data (Cont.).

No.	Attribute	Description
9	Rheumatoid Gout	The Boolean value indicates for rheumatoid Gout disease.
10	Tuberculosis	The Boolean value indicates for tuberculosis disease.
11	Osteoporosis	The Boolean value indicates for osteoporosis.
12	Weight Change	The Boolean value indicates for weight increase or decrease in 3 months.
13	Urinary Incontinence	The Boolean value indicates for urinary incontinence.
14	Vertigo	The Boolean value indicates for vertigo symptom.
15	HIV	The Boolean value indicates for HIV.
16	Liver disease	The Boolean value indicates for liver disease.

Table 3.8 Attributes of Medical Screening Data (Cont.).

No.	Attribute	Description
17	Herpes zoster or Psoriasis	The Boolean value indicates for herpes zoster or Psoriasis.
18	SLE	The Boolean value indicates for systemic lupus erythematosus.
19	Depressive	The Boolean value indicates for depression symptom.
20	Pregnant	The Boolean value indicates for pregnancy status.
21	Kidney	The Boolean value indicates for kidney disease.
22	Family Cancer	The Boolean value indicates for the family members having been cancer.

Table 3.8 Attributes of Medical Screening Data (Cont.).

No.	Attribute	Description
23	Family Heart Disease	The Boolean value indicates for the family members having been heart disease.
24	Family Diabetes	The Boolean value indicates for the family members having been diabetes.
25	Family Stroke	The Boolean value indicates for the family members having been stroke.
26	Family Heredity	The Boolean value indicates for the family members having been heredity.
27	Bleed	The Boolean value indicates for hemophilia.

Table 3.8 Attributes of Medical Screening Data (Cont.).

No.	Attribute	Description
28	Muscle	The Boolean value indicates for problem coordination of muscle.
29	Loss Balance	The Boolean value indicates for loss of balance.

3.10 Data Grouping

In data collection step, some of them are incomplete by some phenomenon such as missing values, noisy data, and erroneous. Thus, they are grouped and arranged as ordinal scheme as shown in Table 3.9.

Table 3.9 Attribute description after data preprocessing process.

No.	Attribute	Description	Values
1	Sex	Gender of sample	Male = 1 Female = 2
2	Age	Age of sample	Integer
3	Province	Province live of sample	Capital = 1 Circumference = 2 Country = 3
4	Marital status	Marital status of sample	Single = 1 Cohabit = 2

Table 3.9 Attribute description after data preprocessing process (Cont.).

No.	Attribute	Description	Values
5	Education	Education of sample	Primary = 1 High school = 2 Vocational = 3 Diploma = 4 Bachelor = 5 Master = 6 PhD. = 7
6	Occupation	Occupation of sample	Government = 1 Merchant = 2 Farmers = 3 Steward = 4

3.11 Modeling

At this stage, the dataset are separated in to three: demographic data, medical screening data, and their integration. Those three datasets are used to create model by three algorithms: Neural Network, Decision tree and Naïve Bayes. To set up the experiment, the neural network is heuristically tuned the parameters which are: number of hidden layer (zero, input nodes equality, output nodes equalities, averaging between input nodes and output nodes, and their summation). The learning rate is varied between 0.2 and 0.5. The momentum is varied between 0.1 and 0.2. The decay value is chosen between true and false. Finally, the training time are chosen from 500 or 10,000 times.

3.12 Evaluation

The evaluation process is performed by 10-Fold Cross Validation. Furthermore, factors analysis for comparison with previous research is also proposed.

3.13 Deployment

Sample application is also developed based on the best selected model in deployment phase. The result will be illustrated in the next chapter.

3.14 Research timeline

The research timeline is shown in Table 3.10.

Table 3.10 Research schedule.

Sequence	Tasks	Data	Timeline: 2015											Timeline: 2016			
			FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	JAN	FEB	MAR	APR
1	Business Understanding and define problem																
2	Study related researches and methodology																
3	Data Selection	Demographic data															
4	Data collection and data preprocessing																
5	Data Mining Tasks and model evaluation																

Table 3.10 Research schedule (Cont.).

Sequence	Tasks	Data	Timeline: 2015											Timeline: 2016			
			FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	JAN	FEB	MAR	APR
6	Research conclusion								■	■							
7	Paper Submission									■							
8	Proposal									■							
9	Conference											■					
10	Data Selection	Demographic data + Medical Screening data											■	■	■		
11	Data collection and data preprocessing										■	■	■	■	■		
12	Data Mining Tasks and model evaluation													■	■		
13	Research conclusion													■	■		
14	Documentation									■	■	■	■	■	■	■	■

This chapter described the overall research methodology. The preliminary results from the beginning stage through the data collection phase are demonstrated. The results of model building, evaluation, factors analysis, deployment, and additional discussion will be explored in the next chapter.

CHAPTER IV
RESULTS AND DISCUSSION

The experimental results of the study "Stroke Risk Prediction Model based on Demographic and Medical Screening Data" are compared between three classification methods: Decision tree, Naïve bays, and Neural Network and three dataset: demographic data, medical screening data, and integrated data. The results are then discussed in accuracy, false-positive rate, false-negative rate, and area under ROC curve (AUC). Additionally, the factors analysis is also provided.

4.1 Experimental Results and Discussion

Experimental results are shown in Table 4.1. As introduced, the results are compared between three classification algorithms and three dataset. Three patients dataset are demographic data, medical screening data, and integrated data. Thee classification algorithms are the decision tree with C4.5 algorithm, Naïve Bayes, and Artificial Neural Network with multi-layer structure and backpropagation algorithm.

Since the neural network method consists of several tunable parameters to achieve the optimal model, those parameters in our considerations are amount of hidden nodes on single hidden layers, learning rate, and momentum. In this research, the tuning process was done by heuristic procedure. An amount of hidden nodes is tuned by selection of following values: zero (*0*), number of input nodes equality (*i*), number of output nodes equality (*o*), summation of *i* and *o*, and average of *i* and *o*. The learning rate is varied between 0.2 and 0.5. The momentum is varied between 0.1 and 0.2. The process of neural network tuning procedure of demographic data, medical screening data, and integrated data are shown in Figure 4.1 to 4.3 respectively. After the experiments are complete, the primary model selection criteria is accuracy and AUC. The secondary criteria is FN rate and FP rate which will be discussed later.

Table 4.1 Comparative Experimental Results.

Method	**Data**	**Accuracy**	**FP rate**	**FN rate**	**AUC**
Decision Tree	Demographic Data	0.73	0.14	0.52	0.72
	Medical Screening Data	0.8	0.11	0.36	0.8
	Integrated Data	0.82	0.11	0.33	0.8
Naïve Bayes	Demographic Data	0.74	0.21	0.37	0.79
	Medical Screening Data	0.79	017	0.28	0.83
	Integrated Data	0.8	0.19	0.22	0.86
Neural Network *Input layer = 6* *hidden layer = 2* *output layer = 2* *learning rate = 0.4* *momentum = 0.1*	Demographic Data	0.77	0.16	0.37	0.8
Neural Network *Input layer = 29* *hidden layer = 2* *output layer = 2* *learning rate = 0.1* *momentum = 0.2*	Medical Screening Data	0.85	0.08	0.31	0.88
Neural Network *Input layer = 35* *hidden layer = 37* *output layer = 2* *learning rate = 0.5* *momentum = 0.2*	Integrated Data	0.84	0.12	0.25	0.9

hiddenlayer	a	i	o	t	0	a	i	o	t	0	a	i	o	t	0
learningrate	0.2	0.2	0.2	0.2	0.2	0.3	0.3	0.3	0.3	0.3	0.4	0.4	0.4	0.4	0.4
momentum	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
Decay	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Training Time	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500
Demo	131	132	133	134	135	136	137	138	139	140	141	142	143	144	145
Accuracy	0.75	0.75	0.76	0.76	0.76	0.76	0.76	0.76	0.76	0.76	0.76	0.76	0.77	0.76	0.76
FP Rate	0.15	0.14	0.16	0.15	0.15	0.16	0.15	0.16	0.15	0.15	0.16	0.16	0.16	0.15	0.15
FN Rate	0.44	0.47	0.42	0.44	0.42	0.41	0.43	0.39	0.4	0.42	0.4	0.4	0.37	0.4	0.42
AUC	0.8	0.79	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8
Screen															
Accuracy	0.84	0.84	0.84	0.83	0.83	0.84	0.84	0.84	0.83	0.83	0.83	0.84	0.83	0.83	0.83
FP Rate	0.09	0.09	0.1	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.08
FN Rate	0.3	0.31	0.3	0.31	0.32	0.31	0.31	0.31	0.32	0.34	0.33	0.31	0.33	0.32	0.33
AUC	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.87	0.88	0.88	0.88	0.88	0.87	0.87	0.88
Demo+Screen															
Accuracy	0.84	0.84	0.84	0.84	0.84	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.82	0.83	0.83
FP Rate	0.1	0.1	0.11	0.1	0.1	0.11	0.11	0.13	0.11	0.11	0.12	0.11	0.13	0.12	0.12
FN Rate	0.27	0.27	0.26	0.28	0.28	0.28	0.27	0.26	0.27	0.29	0.28	0.27	0.29	0.27	0.29
AUC	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.9	0.9	0.9	0.9	0.9

Figure 4.1 Result of MLP and selection (1).

hiddenlayer	a	i	o	t	0	a	i	o	t	0	a	i	o	t	0
learningrate	0.5	0.5	0.5	0.5	0.5	0.1	0.1	0.1	0.1	0.1	0.2	0.2	0.2	0.2	0.2
momentum	0.1	0.1	0.1	0.1	0.1	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2
Decay	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Training Time	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500
Demo	146	147	148	149	150	151	152	153	154	155	156	157	158	159	160
Accuracy	0.77	0.77	0.77	0.76	0.76	0.72	0.7	0.72	0.72	0.76	0.75	0.76	0.76	0.75	0.76
FP Rate	0.16	0.15	0.16	0.16	0.15	0.09	0.06	0.09	0.09	0.15	0.16	0.15	0.15	0.16	0.15
FN Rate	0.37	0.39	0.37	0.39	0.41	0.65	0.76	0.65	0.65	0.42	0.44	0.44	0.42	0.43	0.41
AUC	0.8	0.8	0.8	0.8	0.8	0.78	0.77	0.78	0.78	0.8	0.8	0.8	0.8	0.8	0.8
Screen															
Accuracy	0.83	0.82	0.84	0.83	0.83	0.84	0.83	0.85	0.84	0.84	0.84	0.84	0.84	0.84	0.84
FP Rate	0.09	0.1	0.08	0.09	0.08	0.08	0.09	0.08	0.09	0.09	0.09	0.09	0.09	0.09	0.09
FN Rate	0.33	0.33	0.32	0.33	0.34	0.31	0.33	0.31	0.32	0.31	0.31	0.32	0.3	0.31	0.32
AUC	0.87	0.87	0.87	0.87	0.87	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88
Demo+Screen															
Accuracy	0.82	0.83	0.82	0.83	0.83	0.84	0.83	0.84	0.84	0.84	0.84	0.83	0.84	0.83	0.84
FP Rate	0.13	0.12	0.13	0.12	0.12	0.09	0.1	0.1	0.1	0.1	0.11	0.11	0.11	0.11	0.1
FN Rate	0.28	0.27	0.3	0.27	0.29	0.29	0.31	0.29	0.3	0.3	0.28	0.29	0.27	0.29	0.29
AUC	0.9	0.9	0.9	0.9	0.9	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.91

Figure 4.2 Result of MLP and selection (2).

hiddenlayer	a	i	o	t	0	a	i	o	t	0	a	i	o	t	0
learningrate	0.3	0.3	0.3	0.3	0.3	0.4	0.4	0.4	0.4	0.4	0.5	0.5	0.5	0.5	0.5
momentum	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2
Decay	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Training Time	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500
Demo	161	162	163	164	165	166	167	168	169	170	171	172	173	174	175
Accuracy	0.77	0.76	0.76	0.77	0.76	0.77	0.76	0.77	0.76	0.76	0.77	0.77	0.76	0.77	0.76
FP Rate	0.15	0.15	0.16	0.15	0.15	0.15	0.15	0.16	0.15	0.15	0.16	0.16	0.17	0.16	0.15
FN Rate	0.39	0.41	0.39	0.39	0.42	0.39	0.4	0.37	0.4	0.42	0.37	0.37	0.37	0.37	0.41
AUC	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8
Screen															
Accuracy	0.84	0.84	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.82	0.84	0.83	0.83
FP Rate	0.09	0.09	0.09	0.09	0.08	0.09	0.1	0.09	0.09	0.08	0.09	0.1	0.08	0.1	0.08
FN Rate	0.32	0.31	0.32	0.33	0.34	0.32	0.33	0.33	0.32	0.34	0.34	0.33	0.32	0.33	0.33
AUC	0.88	0.88	0.87	0.88	0.88	0.88	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87
Demo+Screen															
Accuracy	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.82	0.83	0.83	0.83	0.83	0.81	0.84	0.82
FP Rate	0.11	0.11	0.13	0.11	0.11	0.12	0.11	0.13	0.11	0.12	0.12	0.12	0.13	0.12	0.12
FN Rate	0.28	0.27	0.27	0.27	0.29	0.27	0.27	0.29	0.28	0.29	0.28	0.29	0.31	0.25	0.29
AUC	0.91	0.91	0.9	0.91	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9

Figure 4.3 Result of MLP and selection (3).

From the experimental result shown in Table 4.1, it is shown that the Neural Network working with integrated dataset is the best model proven by the 0.84 accuracy and 0.90 AUC. To illustrate and focus the difference between this model and others, the graphical bar charts of accuracy and AUC of all models are shown in Figure 4.4 and 4.5 respectively.

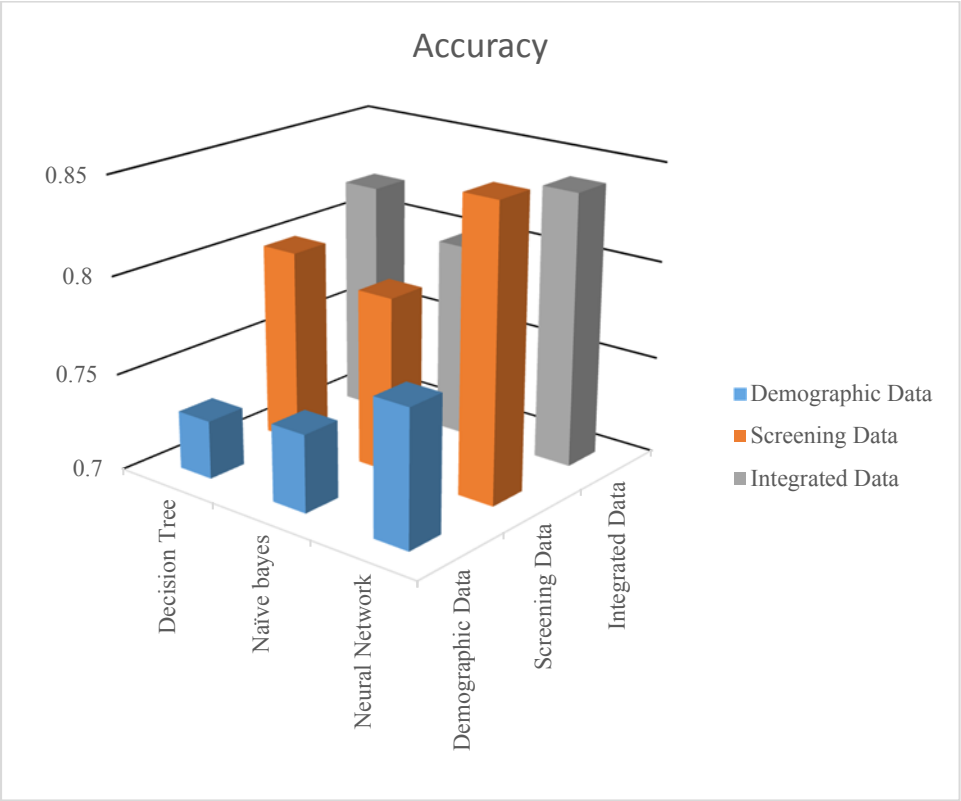


Figure 4.4 Comparison of accuracy for all datasets and algorithms.

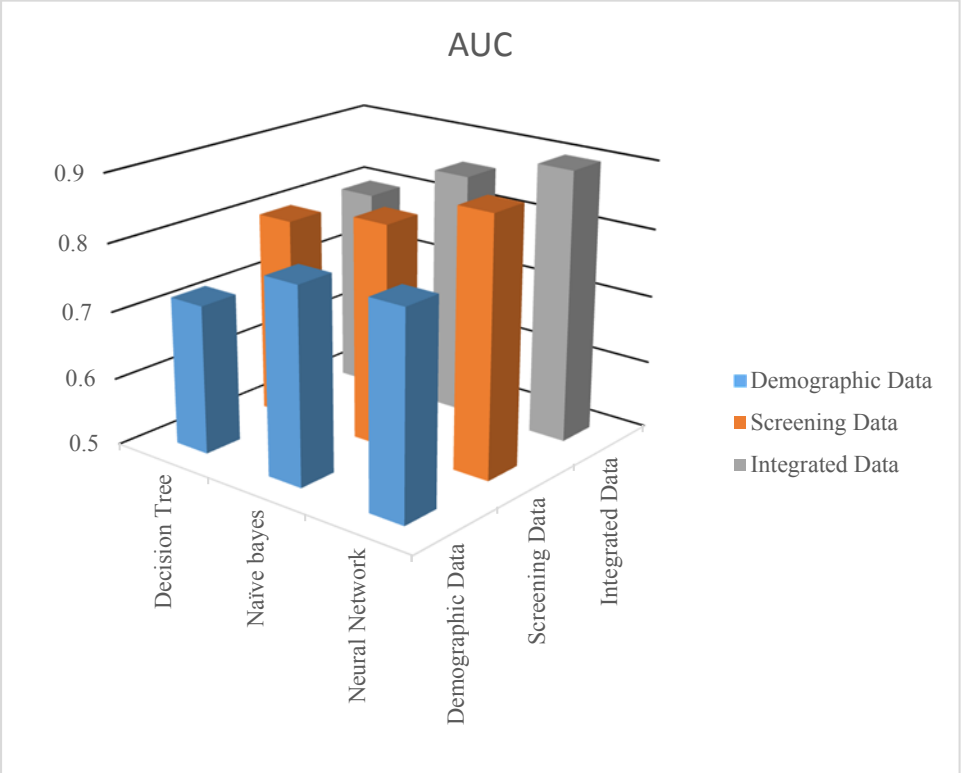


Figure 4.5 Comparison of AUC for all datasets and algorithms.

To focus on the wrong predition, since there are two types of misclassified that should be discussed. The FP rate and FN rate are also considered as illustrated in Figure 4.6 and 4.7. The False-Positive rate (FP rate) is the ratio of number of wrong prediction that patients are stroke and summation of number of wrong prediction that patients are stroke and number of correct prediction in non-stroke patients. To simplify, the FP rate is the ratio between False-Positive number and False-Positive number +True Negative number. Ideally, the FP rate can be high for the prevention benefit. On the other hands, the False-Negative rate (FN rate) is the ratio of number of wrong prediction that patients are non-stroke and summation of number of wrong prediction that patients are non-stroke and number of correct prediction in stroke patients. To simplify, the FN rate is the ratio between False-Negative number and False-Negative number + True Positive number. Ideally and especially in medical field, the FN rate should be a low value for the misdiagnostic benefit which should be pessimistic. However, those values chainly affect to the accuracy.

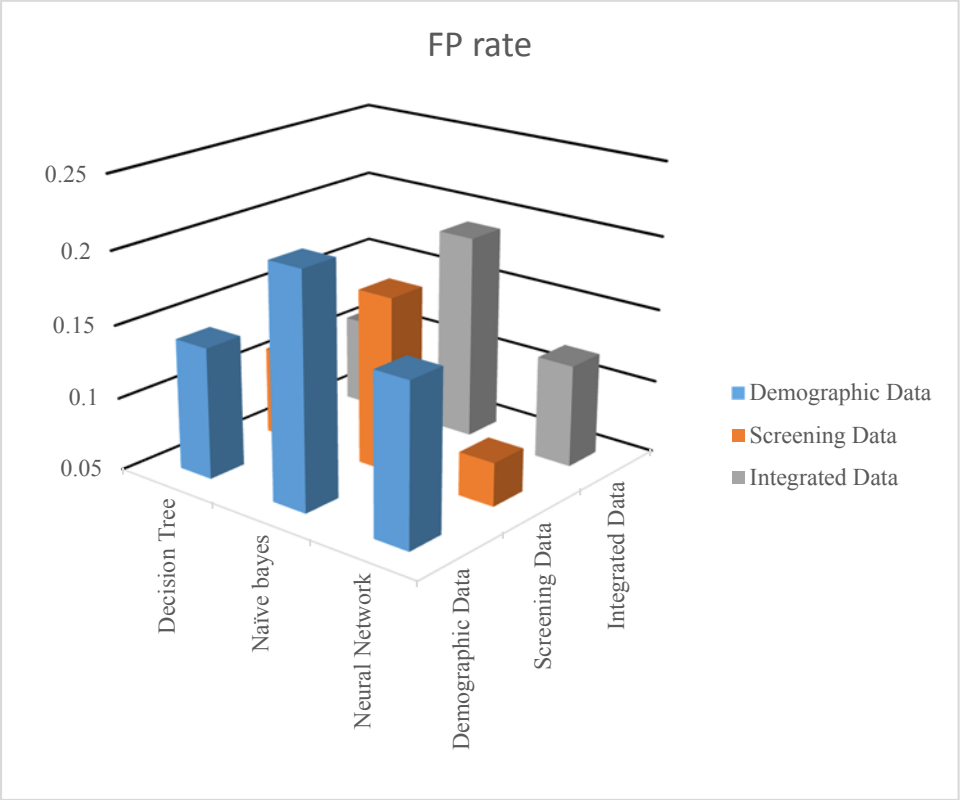


Figure 4.6 Comparison of FP rate for all datasets and algorithms.

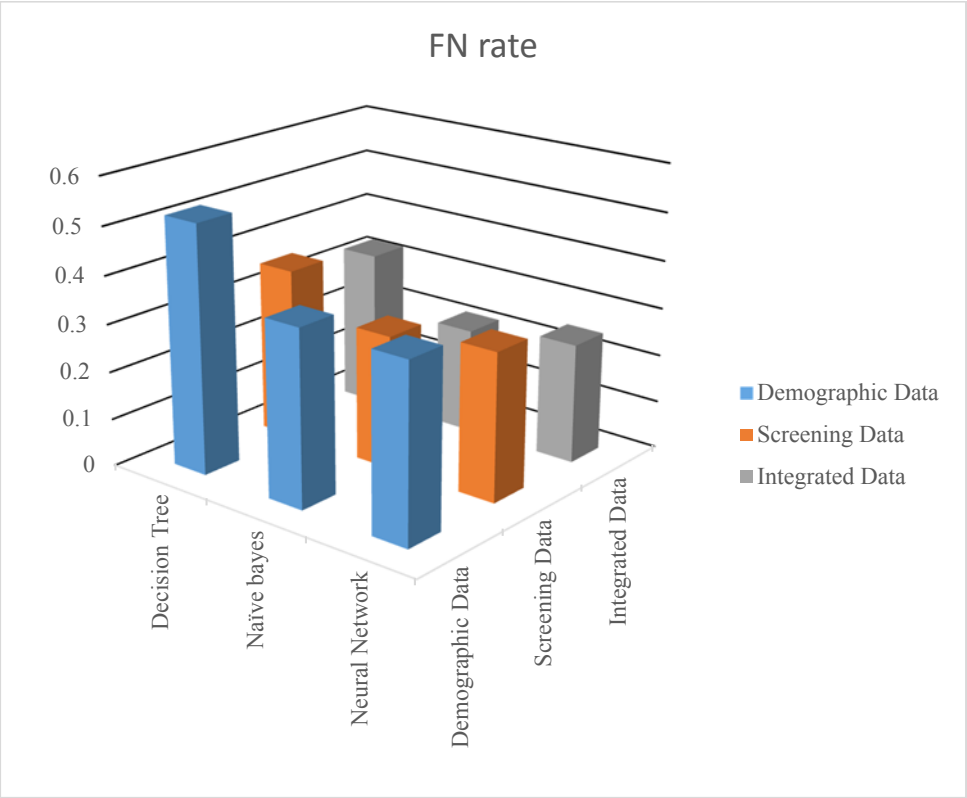


Figure 4.7 Comparison of FN rate for all datasets and algorithms.

The Neural Network with integrated data generates the greatest accuracy and AUC. Moreover since this work relates with life of patient so researcher concern FN rate. Although the Neural Network with integrated data is not the best accuracy, but is the best model because accuracy close the best accuracy and low FN rate.

4.2 Factors Analysis

Additionally, the factor analysis is proposed to explore the risk factors compared with the existing research. The Decision Tree model is chosen to exemplify since its structure is interpretable. The decision tree of each dataset will be explored and analyzed below. Firstly, the decision tree of demographic data can be found that the age is the main factor that age more than 62 years old Female and Education less than high school or equal chance is stroke. Otherwise, an education higher than high school is non-stroke as shown in Figure 4.8.

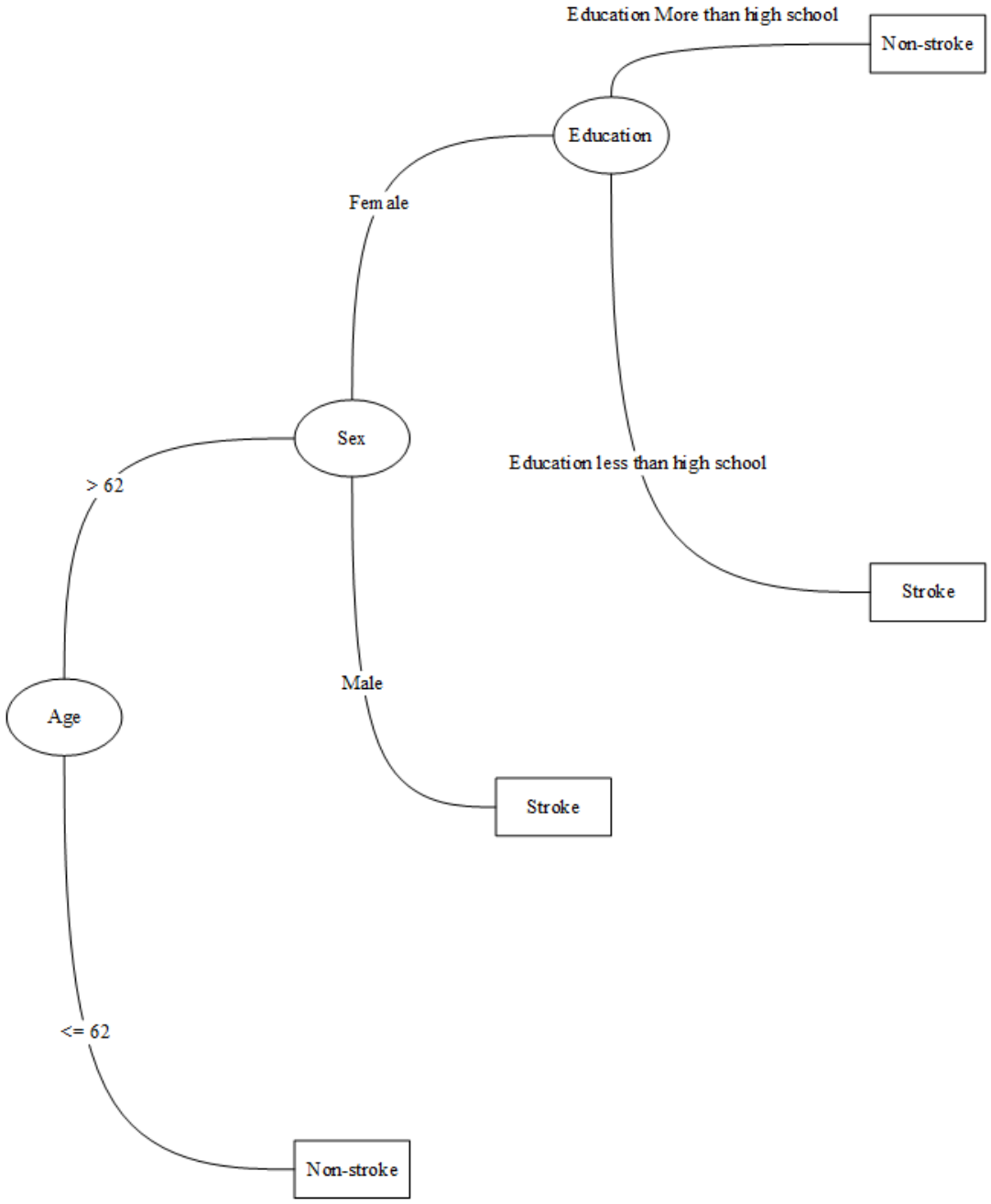


Figure 4.8 Decision Tree of Demographic Data.

As shown in Figure 4.9, the Decision Tree of the medical screening data can be found that the risk factor is hypertension, the stroke-family members, and balance losing is stroke while non-hypertension, non-diabetes and non-hyperlipidemia patients is classified as non-stroke.

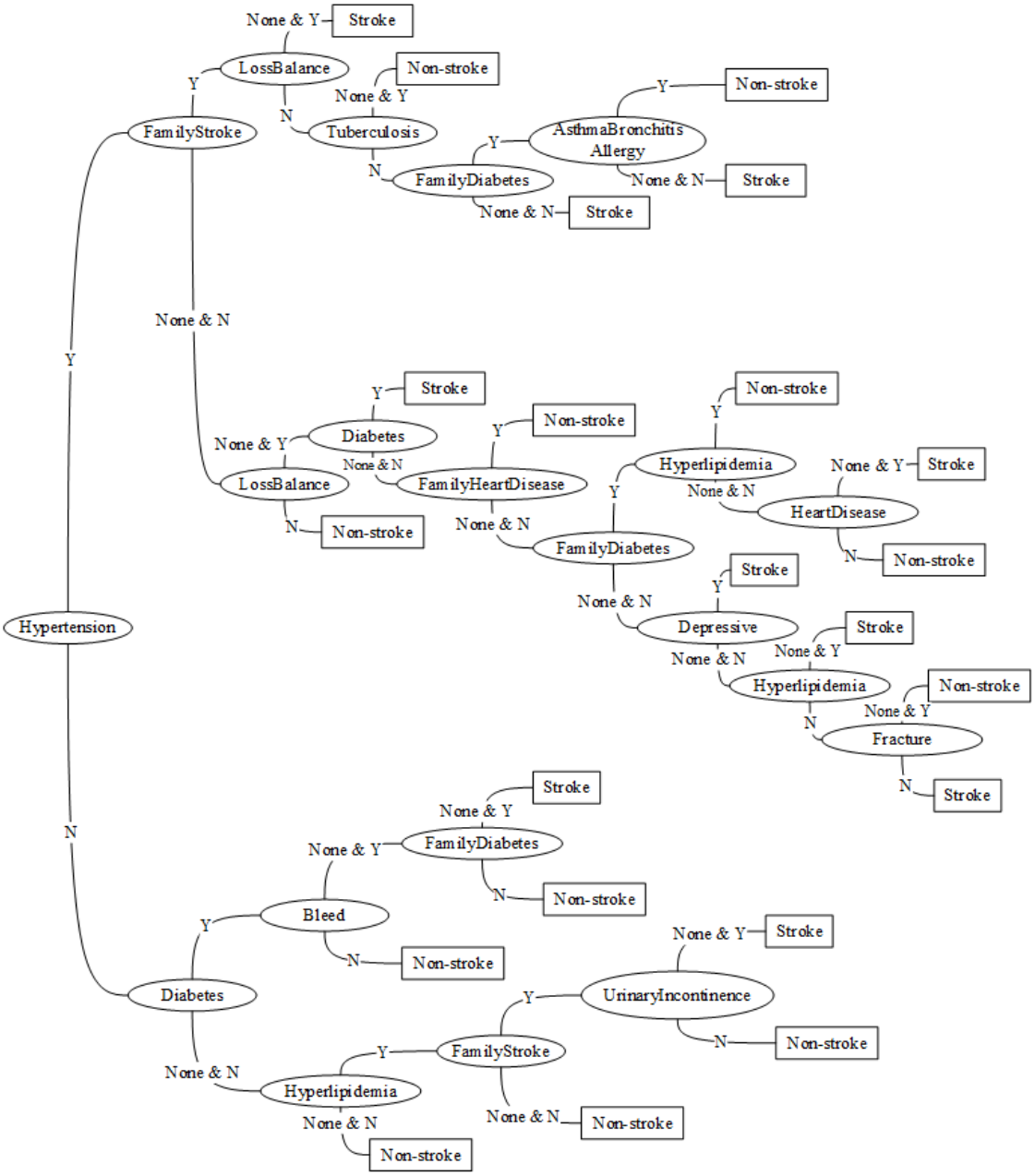


Figure 4.9 Decision Tree of Medical Screening Data.

Finally, the Decision Tree of integrated dataset can be found that the patients who are non-hypertension, non-diabetes, below 68 years old, hyperlipidemia, and male will face a chance of stroke while those who diabetes, bleed, and non-diabetes family members may be non-stroke. The Decision Tree is shown in Figure 4.10.

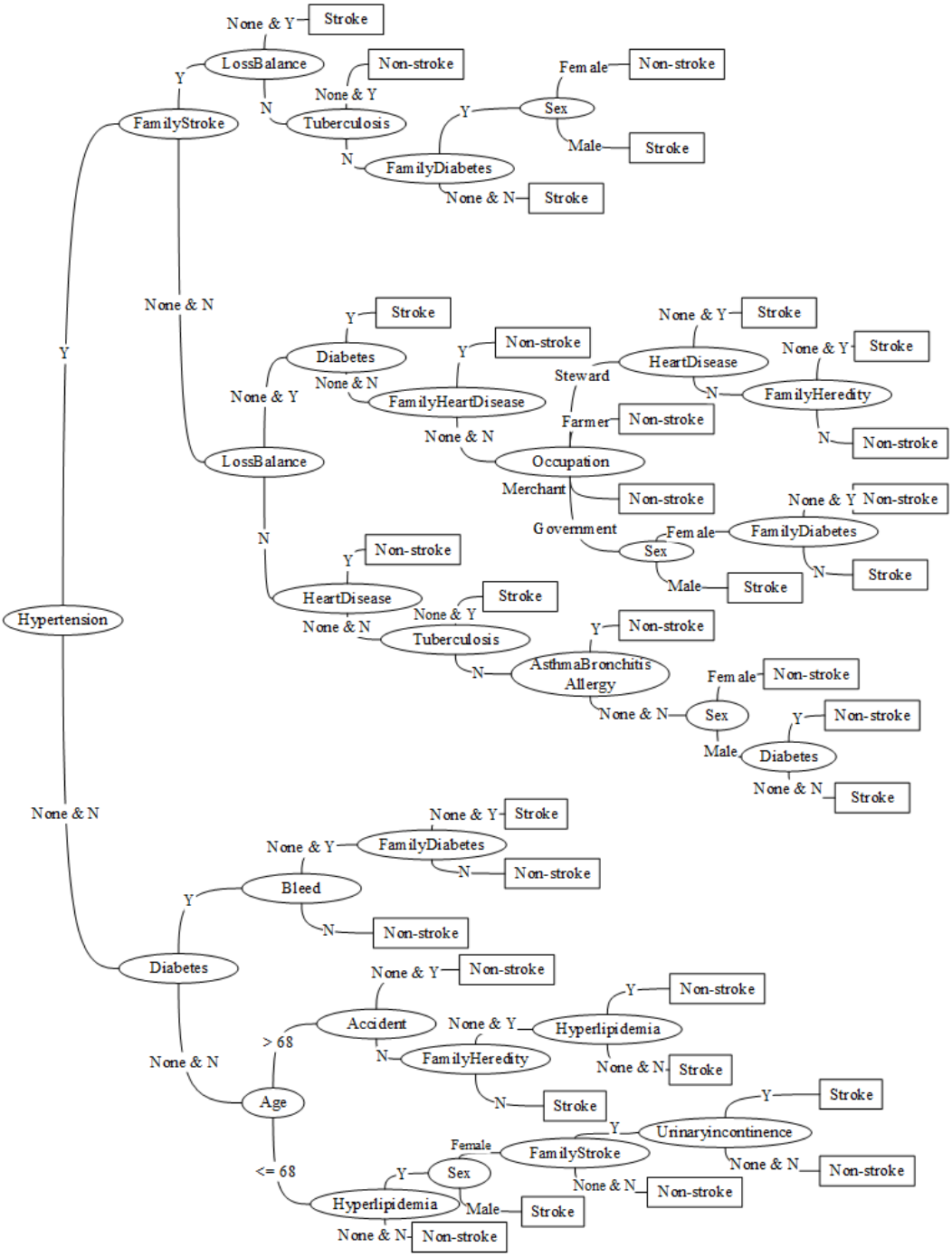


Figure 4.10 Decision Tree of Integrated Data.

The result of integrated data consists of factors appeared and lose in from demographic data and medical screening data. For example, the occupation factor has been added. Some factors may be was replaced such as hyperlipidemia which is replaced by age. That is age is more important factor than hyperlipidemia. Moreover bleeding, hemophilia in the third level of decision tree, and losing balance in the third level of

decision tree have been never mentioned in the previous research of stroke risk factors (Dutsadeevettakul, 2013) (Boysen *et al.*, 1988). Thus, apart from the prediction model, this research also proposes the methodology for risk factors discovering. However, since the bleeding factor is appeared in the oldest form, the results may indiscrepant between unknown and known values.

4.3 Deployment

After modeling and evaluation is complete, the final step is deployment. It brings the best results from modeling to create a tool from in spread sheet and illustrated some input which is simulated as new patient information as shown in Figure 4.11 and Figure 4.12.

From Figure 4.11, the tool can predict that a male patient with age equal seventy years old, living in suburb, cohabit marital status, hypertension, diabetes, hyperlipidemia, vertigo, and family members appear to be cancer and diabetes, would be predicted as stroke. Another example is shown in Figure 4.12. The tool can predict that female patients with age equal forty years old, living in suburb, single marital status, non-hypertension, non-diabetes, be cancer, be herpes zoster, be SLE and be losing of balance, would be diagnosed as non-stroke.

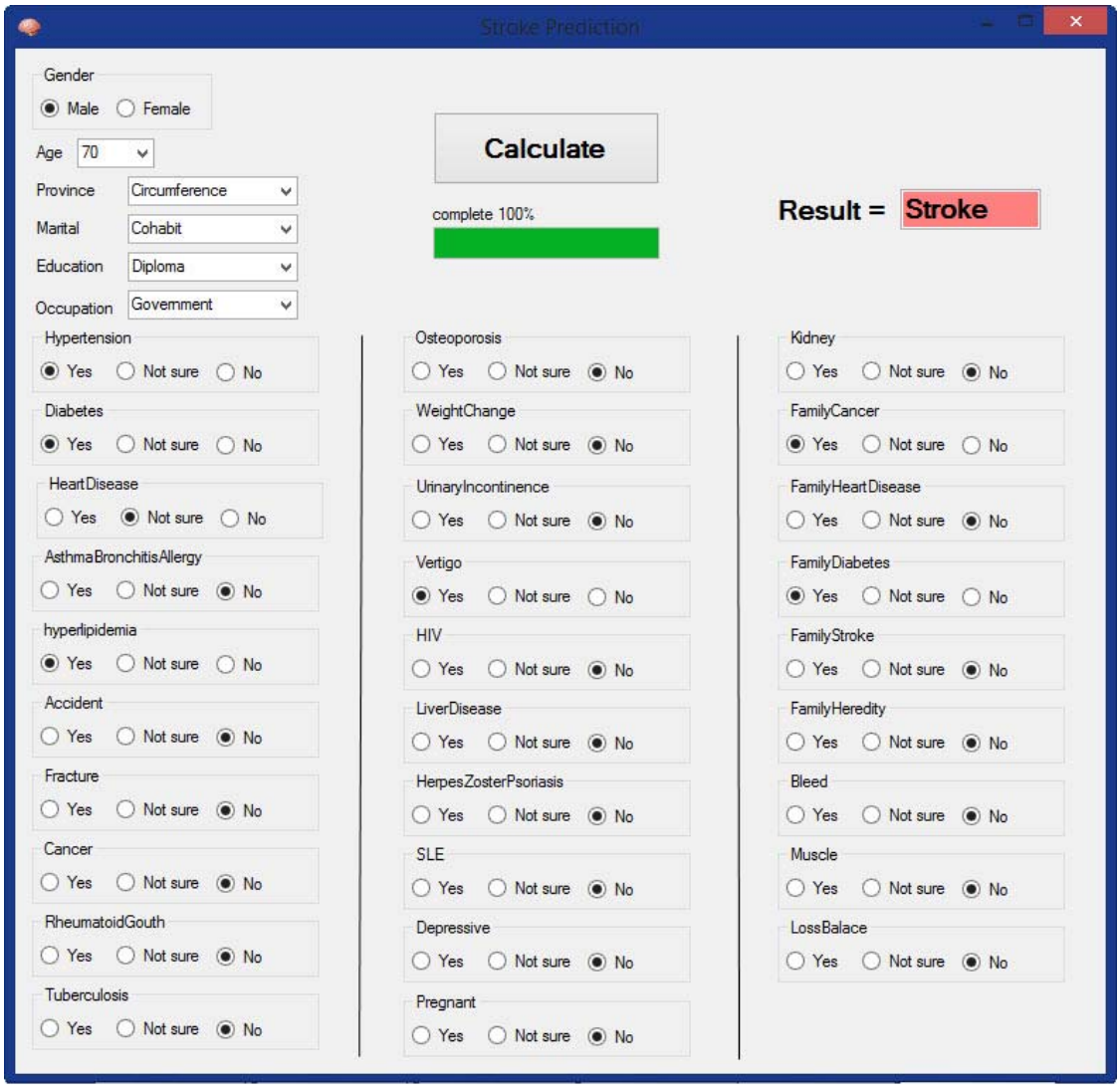


Figure 4.11 A sample data input for stroke prediction.

Figure 4.12 A sample data input for non-stroke prediction.

This chapter presents the experimental result, discussion, factors analysis, and deployment. It is found that the most appropriate prediction model is the neural network working with integrated dataset. Factors analysis is also investigated and compared with previous risk factors finding. Finally, a tool based on the best model is demonstrated. To conclude all works, the conclusion will be shown in the next chapter.

CHAPTER V
CONCLUSION

This chapter concludes over all research and then states the limitation and the opportunity in future work.

5.1 Conclusion

This research proposed the stroke risk prediction model using demographic data and medical screening data applied by three classification algorithms which are Neural Network, Decision Tree, and Naïve Bayes. This research was approved from IRB and data owner in order to gather and collect demographic data. After data acquiring, researcher processed the resampling to reduce the data ratio between stroke and non-stroke patients amount to be 1:2. Then, researcher collected the data from medical screening files and perform post-resampling due to the imbalance data caused by the lost medical files. Consequently, the medical screening data were integrated into single data format since there have been several versions of forms. Then, the data were grouped before creating model by three methods with 10-fold cross validation method. The deployment is proposed by applied the best model to the spread sheet application. The best model from the experiment is the Neural Network with integrated data by accuracy 0.84, FP rate 0.12, FN rate 0.25 and AUC 0.9. The accuracy and AUC are the primary criteria of model selection while the FP rate and FN rate are the validating criteria in medical field research.

Thus, the benefit of this research is to find the best stroke risk prediction model and to discover novel risk factors. The result of integrated data, using the best decision tree due to its interpretable result, consists of factors appeared and lose in single demographic data and medical screening data. Some factors, which are hemophilia and losing balance, are superior discovered to previous research.

5.2 Limitation

This research develops a stroke risk prediction model based on the data collected from Faculty of Physical Therapy during 2012 - 2015. Since the clinic is located near to the capital city, bias in data source may be hidden in the research due to the behavior of patients. This may be furtherly arguable that the proposed model can be represent for the whole citizen in Thailand. Additionally, since to medical screening data has been directly collected from the patients, the linguistic bias may be attached in the research because some patients may not clearly understand the definition of the medical terms stated in the form. Furthermore in technical aspect, since the resampling process is relied on the hypothesis that the distribution of original data conforms with the normal distribution as evidenced by using arithmetic mean as a considered criteria. The more insight experiment in data distribution should be attached in the research.

5.3 Future work

To extend and enhance the research to be more realistic and practical, the researcher offers: 1) to expand the data source in order to discover a model that representable for whole citizen of the country, 2) to consider some more complicate classification algorithm to improve the prediction model, 3) the deeper data exploration and analysis for the benefit of the best resampling method selection, and 4) since the discovered primary factors are consistent with the prior research, the minor factors scoping by pre-eliminating the well-known factors is also challenge.

REFERENCES

Adamson, J., Beswick, A., & Ebrahim, S. (2004). Is stroke the most common cause of disability? *Journal of Stroke and Cerebrovascular Diseases*, 171-177.

Berry, M. J. A., & Linoff, G. S. (1997). *Data Mining Techniques: For Marketing, Sales, and Customer Support*. Chichester: Wiley-Blackwell.

Bradley, A. P. (1997). *The use of the Area under the ROC Curve in the Evaluation of Machine Learning Algorithms*. Canada.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0 Step-by-step data mining guide*.

Duda, R. O., Hart, P. E., & Stork, D. G. (1973). *Pattern Classification* (1 ed.): Wiley-Blackwell.

Giudici, P. (2003). *Applied Data Mining Statistical Methods for Business and Industry*. Chichester: Wiley-Blackwell.

Han, J., & Kamber, M. (2006). *Data Mining: Concepts and Techniques* (2 ed.). USA: Elsevier Inc.

Hanchaiphiboolkul, S., Poungvarin, N., Nidhinandana, S., Suwanwela, N. C., Puthkhao, P., Towanabut, S., Tantirittisak, T., Suwantamee, J., Samsen, M. (2011). Prevalence of Stroke and Stroke Risk Factors in Thailand: Thai Epidemiologic Stroke (TES) Study. *J Med Assoc Thai, 94*, 427-436.

Kohavi, R. (1995). *A study of cross-validation and bootstrap for accuracy estimation and model selection.* Paper presented at the 14th international joint conference on Artificial

Kohavi, R., & Provost, F. (2001). *Applications of Data Mining to Electronic Commerce*: Springer.

Laidler, P. (1994). *Stroke Rehabilitaion* (1 ed.). London: Chapman & Hall.

Leal, J., Luengo-Ferna´ndez, R. n., Gray, A., Petersen, S., & Rayner, M. (2006). Economic burden of cardiovascular diseases in the enlarged European Union. *European Heart Journal*. doi: 10.1093/eurheartj/ehi733

Longadge, R., Dongre, S. S., & Malik, L. (2013). Class Imbalance Problem in Data Mining: Review. *International Journal of Computer Science and Network (IJCSN), 2*(1).

Macqueen, J. (1967). *Some methods for classification and analysis of mulivariate observations*. University of California, Los Angeles.

Occupational Therapy and Stroke. (2010). (J. Edmans Ed. 2nd ed. ed.). Chichester, West Sussex, U.K.: Wiley-Blackwell.

Office, N. A. (2005). Department of Health: Progress in improving stroke care. Retrieved 20 Jan, 2015, from https://www.nao.org.uk

Poungvarin, N. (2001). *Stroke* (2 ed.). Bangkok: Siriraj hospital.

Prati, R. C., Batista, G. E. A. P. A., & Monard, M. C. *Data mining with imbalanced class distributions: concepts and methods*. Paper presented at the Indian International Conference on Artificial Intelligence (IICAI-09), India.

Roiger, R., & Geatz, M. (2003). *Data Mining: A Tutorial Based Primer*. United States of America: Addison Wesley.

Roquer, J., Campello, A. R., & Gomis, M. (2003). Sex differences in first-ever acute stroke. 1581-1585. doi: 10.1161/01.STR.0000078562.82918.F6

Sinsomboonthong, S. (2015). *Data Mining* (1 ed.). Bangkok: Jamjuree.

Spence, D., & Barnett, H. J. (2012). *Stroke prevention, treatment, and rehabilitation*. New York: McGraw-Hill Medical.

Sudha, A., Gayathri, P., & Jaisankar, N. (2012). Effective Analysis and Predictive Model of Stroke Disease using Classification Methods. *International Journal of Computer Applications (0975 – 8887), 43*, 26-31.

Townsend, N., Wickramasinghe, K., Bhatnagar, P., Smolina, K., Nichols, M., Leal, J., Fernandez, R., Rayner, M. (2005). *Coronary heart disease statistics* (2005 ed.). London: the British Heart Foundation.

WHO. (1994). International Classification of Diseases (ICD). Retrieved 20 Jan, 2015, from http://www.who.int/classifications/icd/en/

WHO. (2015). *World Health Statistics*. Geneva: WHO.

Wolfe, C. D. A. (2000). The impact of stroke. *British Medical Bulletin*, 275-286.

Zhu, X., & Davidson, I. (2007). *Knowledge discovery and data mining: challenges and realities. Premier reference source. Information Science Reference* (I. Davidson Ed.): IGI Global.

APPENDICES

APPENDIX A
MEDICAL SCREENING FORMS

แบบสอบถาม

วัตถุประสงค์ เพื่อให้เกิดความเข้าใจในภาวะสุขภาพปัจจุบันของผู้ป่วย โปรดตอบแบบสอบถามโดยสมบูรณ์และสามารถซักถามข้อสงสัยจากนักกายภาพบำบัดได้ (แบบสอบถามนี้จะถูกบันทึกเป็นประวัติการเจ็บป่วยส่วนตัวของท่าน)

ชื่อ [redacted] นามสกุล [redacted]

อายุ 48 ปี เพศ [illegible] ส่วนสูง 160 เซนติเมตร น้ำหนัก 60 กิโลกรัม

ประวัติอาการเจ็บป่วย : (ให้ใส่เครื่องหมาย ✓ ในช่องเคยหรือไม่เคย)

ท่านเคยมีประวัติการเจ็บป่วยเหล่านี้หรือไม่	เคย	ไม่เคย	ข้อมูลบันทึกเพิ่มเติม (เฉพาะนักกายภาพบำบัด)
1. ความดันสูง		✓	
2. โรคเบาหวาน		✓	
3. โรคหัวใจ หรือเสียงเต้นหัวใจผิดปกติ	✓		
4. เจ็บหน้าอก (angine) ขณะพัก หรือทำกิจกรรม	✓		
5. โรคระบบทางเดินหายใจ เช่น หืดหอบ, หลอดลมอักเสบเรื้อรัง		✓	
6. โรคภูมิแพ้ หรืออาการแพ้ต่าง ๆ	✓		
7. โรคไต		✓	
. อุบัติเหตุต่าง ๆ	✓		
9. น้ำตาลในเลือดต่ำ		✓	
10. โรคมะเร็ง		✓	
11. อาการบวม หรืออักเสบตามข้อต่าง ๆ		✓	
12. วัณโรคปอด หรือวัณโรคกระดูก		✓	
13. โรคกระดูกพรุน		✓	
14. ภาวะกระดูกหัก หรือร้าว	✓		
15. น้ำหนักตัวเพิ่ม หรือลดลงมากผิดปกติ		✓	
16. เลือดไหลไม่หยุด	✓		
17. ปัญหาการทำงานประสานกันของกล้ามเนื้อ		✓	
18. การสูญเสียการทรงตัว หรือปัญหาล้มบ่อย ๆ	✓		
19. ปัสสาวะบ่อย, ปัสสาวะไม่ออก หรือกลั้นปัสสาวะไม่ได้	✓		

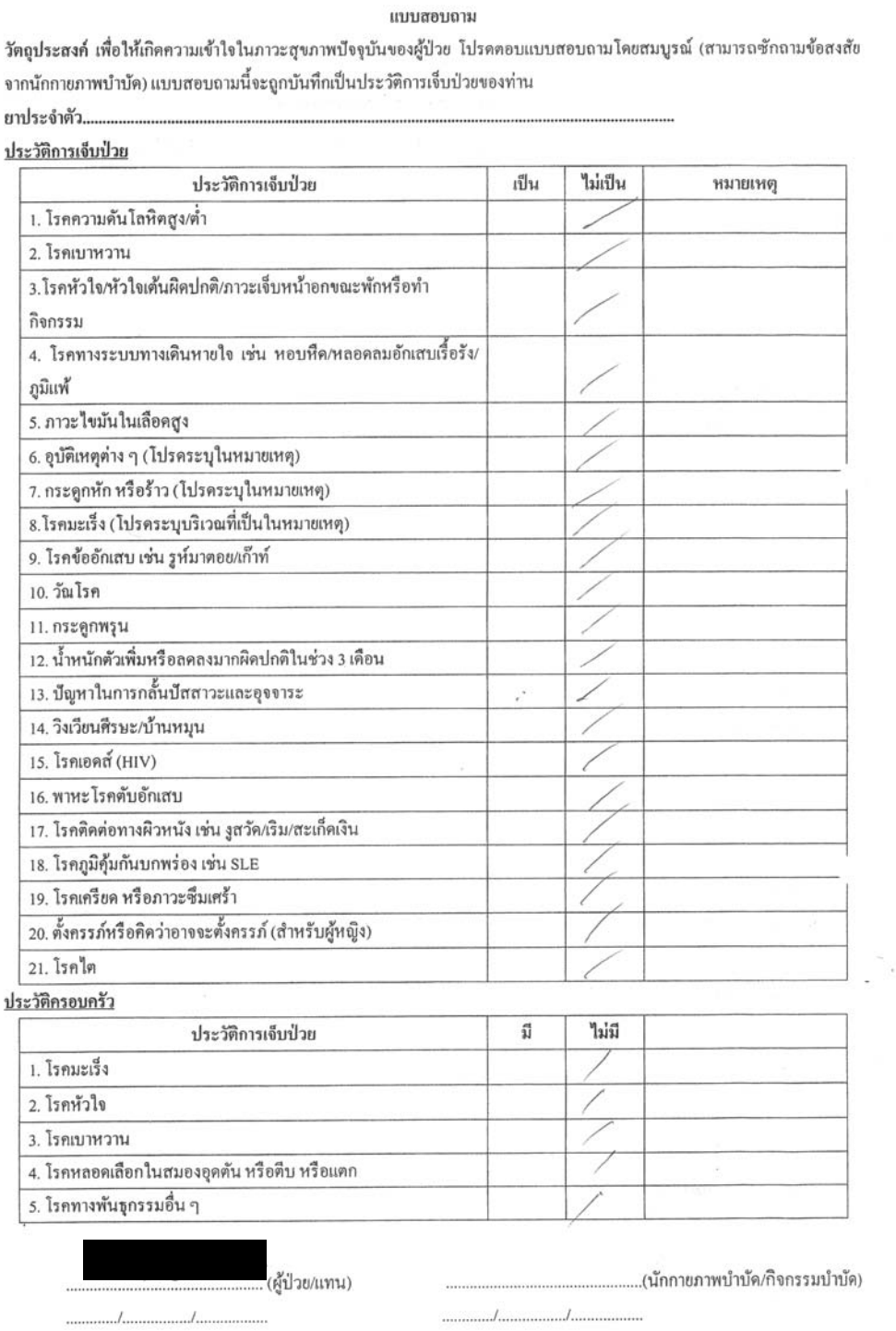
ประวัติครอบครัว

ท่านมีบุคคลภายในครอบครัวที่มีปัญหา หรือเคยเป็นโรคต่าง ๆ เหล่านี้หรือไม่	เคย	ไม่เคย	ข้อมูลบันทึกเพิ่มเติม (เฉพาะนักกายภาพบำบัด)
1.โรคมะเร็ง	✓		
2. โรคหัวใจ	✓		
3. โรคเบาหวาน	✓		
4. โรคหลอดเลือดในสมองอุดตัน หรือแตก	✓		
5. โรคทางพันธุกรรมอื่น ๆ		✓	

ประวัติส่วนตัว : (สำหรับผู้หญิง)

	ใช่	ไม่ใช่	ข้อมูลบันทึกเพิ่มเติม (เฉพาะนักกายภาพบำบัด)
1. ท่านตั้งครรภ์ หรือคิดว่าอาจจะตั้งครรภ์หรือไม่		✓	
2. ขณะนี้ท่านมีอาการเหล่านี้			
- การมาของรอบเดือนไม่สม่ำเสมอ		✓	
- โรคที่มีการอักเสบภายในอุ้มเชิงกราน		✓	
- อาการเยื่อบุมดลูกอักเสบ (Endometriogls)		✓	

Figure A.1 Medical Screening Form during 2008 - 2012.



แบบสอบถาม

วัตถุประสงค์ เพื่อให้เกิดความเข้าใจในภาวะสุขภาพปัจจุบันของผู้ป่วย โปรดตอบแบบสอบถามโดยสมบูรณ์ (สามารถซักถามข้อสงสัยจากนักกายภาพบำบัด) แบบสอบถามนี้จะถูกบันทึกเป็นประวัติการเจ็บป่วยของท่าน

ยาประจำตัว.....

ประวัติการเจ็บป่วย

ประวัติการเจ็บป่วย	เป็น	ไม่เป็น	หมายเหตุ
1. โรคความดันโลหิตสูง/ต่ำ		/	
2. โรคเบาหวาน		/	
3.โรคหัวใจ/หัวใจเต้นผิดปกติ/ภาวะเจ็บหน้าอกขณะพักหรือทำกิจกรรม		/	
4. โรคทางระบบทางเดินหายใจ เช่น หอบหืด/หลอดลมอักเสบเรื้อรัง/ภูมิแพ้		/	
5. ภาวะไขมันในเลือดสูง		/	
6. อุบัติเหตุต่าง ๆ (โปรดระบุในหมายเหตุ)		/	
7. กระดูกหัก หรือร้าว (โปรดระบุในหมายเหตุ)		/	
8.โรคมะเร็ง (โปรดระบุบริเวณที่เป็นในหมายเหตุ)		/	
9. โรคข้ออักเสบ เช่น รูห์มาตอย/เก๊าท์		/	
10. วัณโรค		/	
11. กระดูกพรุน		/	
12. น้ำหนักตัวเพิ่มหรือลดลงมากผิดปกติในช่วง 3 เดือน		/	
13. ปัญหาในการกลั้นปัสสาวะและอุจจาระ		/	
14. วิงเวียนศีรษะ/บ้านหมุน		/	
15. โรคเอดส์ (HIV)		/	
16. พาหะโรคตับอักเสบ		/	
17. โรคติดต่อทางผิวหนัง เช่น งูสวัด/เริม/สะเก็ดเงิน		/	
18. โรคภูมิคุ้มกันบกพร่อง เช่น SLE		/	
19. โรคเครียด หรือภาวะซึมเศร้า		/	
20. ตั้งครรภ์หรือคิดว่าอาจจะตั้งครรภ์ (สำหรับผู้หญิง)		/	
21. โรคไต		/	

ประวัติครอบครัว

ประวัติการเจ็บป่วย	มี	ไม่มี	
1. โรคมะเร็ง		/	
2. โรคหัวใจ		/	
3. โรคเบาหวาน		/	
4. โรคหลอดเลือกในสมองอุดตัน หรือตีบ หรือแตก		/	
5. โรคทางพันธุกรรมอื่น ๆ		/	

[redacted].....(ผู้ป่วย/แทน) .....(นักกายภาพบำบัด/กิจกรรมบำบัด)

...../...../..... ...../...../.....

Figure A.2 Medical Screening Form during 2012 - 2013.

วัตถุประสงค์ในการกรอกข้อมูล : เพื่อให้เกิดความเข้าใจในภาวะสุขภาพปัจจุบันของผู้ป่วยและใช้เป็นข้อมูลช่วยประกอบการทำกายภาพได้อย่างสมบูรณ์ โปรดกรอกข้อมูลให้ครบถ้วน (สามารถซักถามข้อสงสัยใดๆ ได้จากนักกายภาพบำบัดผู้ทำการรักษา) ข้อมูลนี้จะถูกบันทึกประวัติการเจ็บป่วยของท่านไว้

กรุณาใช้เครื่องหมาย ✓ ลงในช่องตามความเป็นจริง

ประวัติการเจ็บป่วยส่วนตัว

ประวัติการเจ็บป่วย	เป็น	ไม่เป็น	ไม่ทราบ	หมายเหตุ
1. โรคความดันโลหิตสูงหรือต่ำ		✓		
2. โรคเบาหวาน		✓		
3. โรคหัวใจ/หัวใจเต้นผิดปกติ/ภาวะเจ็บหน้าอกขณะพักหรือทำกิจกรรม		✓		
4. โรคหอบหืด/หลอดลมอักเสบเรื้อรัง/ภูมิแพ้		✓		
5. โรคเกี่ยวกับภาวะไขมันในเลือดสูง		✓		
6. เกิดอุบัติเหตุต่าง ๆ (โปรดระบุในช่องหมายเหตุ)		✓		
7. กระดูกหักหรือร้าว (โปรดระบุในช่องหมายเหตุ)		✓		
8. โรคมะเร็ง (โปรดระบุประเภทที่เป็นในช่องหมายเหตุ)		✓		
9. โรคข้ออักเสบ เช่น รูห์มาตอยหรือเก๊าท์		✓		
10. โรควัณโรค		✓		
11. โรคกระดูกพรุน		✓		
12. มีภาวะน้ำหนักตัวเพิ่มหรือลดลงมากผิดปกติในช่วง 3 เดือน		✓		
13. มีปัญหาในการกลั้นปัสสาวะและอุจจาระ		✓		
14. มีภาวะการวิงเวียนศีรษะบ่อยครั้ง		✓		
15. โรคเอดส์ (HIV) หรือโรคที่ติดต่อทางเพศอื่นๆ		✓		
16. โรคตับอักเสบหรือเกี่ยวกับตับ		✓		
17. โรคติดต่อทางผิวหนัง เช่น งูสวัด/เริม/สะเก็ดเงิน		✓		
18. โรคภูมิคุ้มกันบกพร่องอื่นๆ เช่น SLE		✓		
19. มีภาวะโรคเครียด , ซึมเศร้าบ่อยครั้ง		✓		
20. มีตั้งครรภ์หรือคิดว่าอาจจะตั้งครรภ์ (สำหรับผู้หญิง) ณ เวลานี้		✓		
21. โรคไตหรือโรคที่เกี่ยวข้อง		✓		
22. อื่นๆ ถ้ามี (เช่น แพ้ยาประเภท ฯลฯ) ระบุ.....				

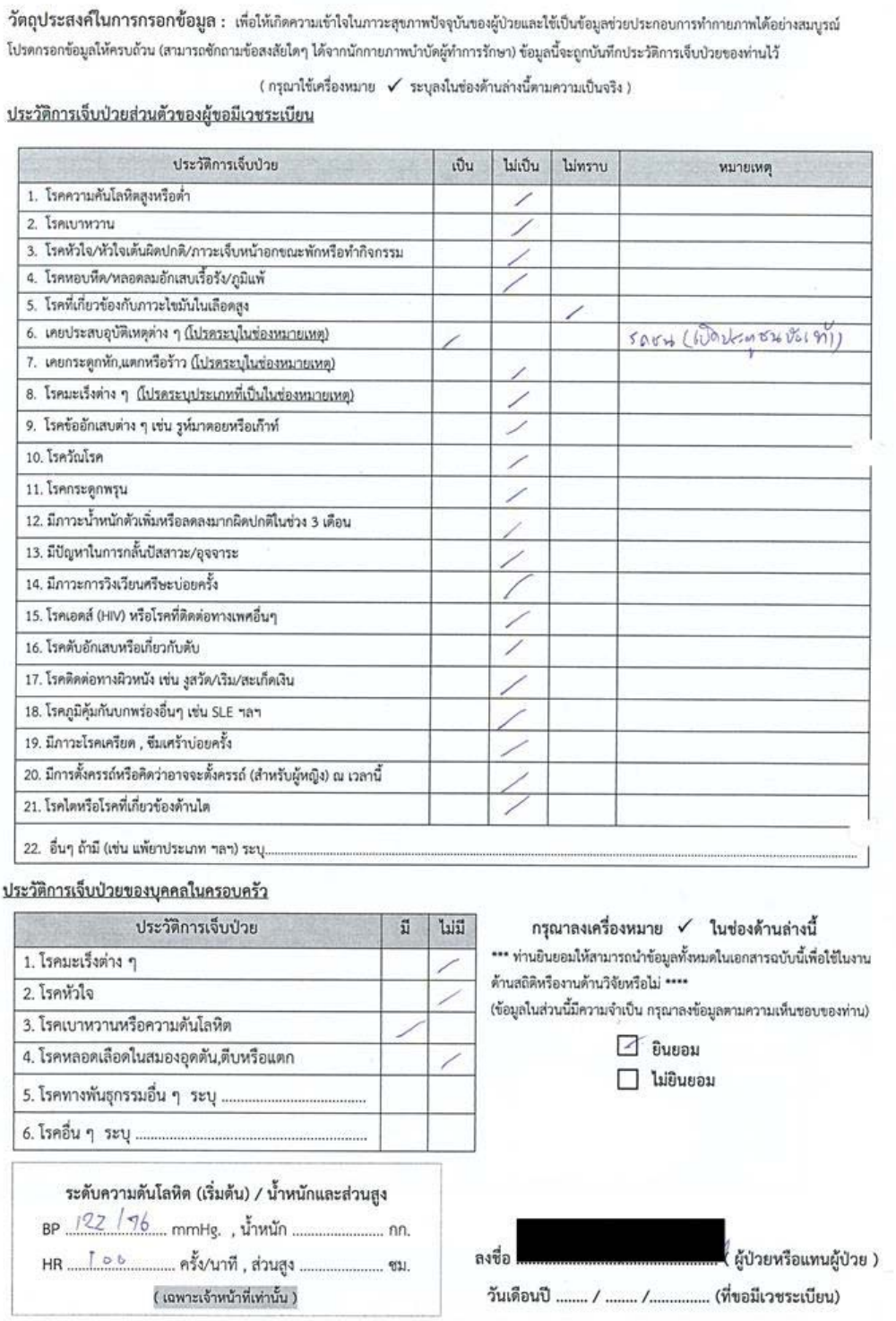
ประวัติการเจ็บป่วยของบุคคลในครอบครัว

ประวัติการเจ็บป่วย	มี	ไม่มี
1. โรคมะเร็งต่าง ๆ		✓
2. โรคหัวใจ		✓
3. โรคเบาหวานหรือความดันโลหิต	✓	
4. โรคหลอดเลือดในสมองอุดตัน ตีบหรือแตก		✓
5. โรคทางพันธุกรรมอื่น ๆ ระบุ		
6. โรคอื่น ๆ ระบุ		

ลงชื่อ [redacted] (ผู้ป่วยหรือแทนผู้ป่วย)

วันเดือนปี 4 ส.ค. 2557

Figure A.3 Medical Screening Form during 2013 - 2015.



วัตถุประสงค์ในการกรอกข้อมูล : เพื่อให้เกิดความเข้าใจในภาวะสุขภาพปัจจุบันของผู้ป่วยและใช้เป็นข้อมูลช่วยประกอบการทำกายภาพได้อย่างสมบูรณ์ โปรดกรอกข้อมูลให้ครบถ้วน (สามารถซักถามข้อสงสัยใดๆ ได้จากนักกายภาพบำบัดผู้ทำการรักษา) ข้อมูลนี้จะถูกบันทึกประวัติการเจ็บป่วยของท่านไว้

(กรุณาใช้เครื่องหมาย ✓ ระบุลงในช่องด้านล่างนี้ตามความเป็นจริง)

ประวัติการเจ็บป่วยส่วนตัวของผู้ขอมีเวชระเบียน

ประวัติการเจ็บป่วย	เป็น	ไม่เป็น	ไม่ทราบ	หมายเหตุ
1. โรคความดันโลหิตสูงหรือต่ำ		✓		
2. โรคเบาหวาน		✓		
3. โรคหัวใจ/หัวใจเต้นผิดปกติ/ภาวะเจ็บหน้าอกขณะพักหรือทำกิจกรรม		✓		
4. โรคหอบหืด/หลอดลมอักเสบเรื้อรัง/ภูมิแพ้		✓		
5. โรคที่เกี่ยวข้องกับภาวะไขมันในเลือดสูง			✓	
6. เคยประสบอุบัติเหตุต่าง ๆ (โปรดระบุในช่องหมายเหตุ)	✓			รถชน (เปิดประทุนขับเที่ย)
7. เคยกระดูกหัก,แตกหรือร้าว (โปรดระบุในช่องหมายเหตุ)		✓		
8. โรคมะเร็งต่าง ๆ (โปรดระบุประเภทที่เป็นในช่องหมายเหตุ)		✓		
9. โรคข้ออักเสบต่าง ๆ เช่น รูห์มาตอยหรือเก๊าท์		✓		
10. โรควัณโรค		✓		
11. โรคกระดูกพรุน		✓		
12. มีภาวะน้ำหนักตัวเพิ่มหรือลดลงมากผิดปกติในช่วง 3 เดือน		✓		
13. มีปัญหาในการกลั้นปัสสาวะ/อุจจาระ		✓		
14. มีภาวะการวิงเวียนศีรษะบ่อยครั้ง		✓		
15. โรคเอดส์ (HIV) หรือโรคที่ติดต่อทางเพศอื่นๆ		✓		
16. โรคตับอักเสบหรือเกี่ยวกับตับ		✓		
17. โรคติดต่อทางผิวหนัง เช่น งูสวัด/เริม/สะเก็ดเงิน		✓		
18. โรคภูมิคุ้มกันบกพร่องอื่นๆ เช่น SLE ฯลฯ		✓		
19. มีภาวะโรคเครียด , ซึมเศร้าบ่อยครั้ง		✓		
20. มีการตั้งครรภ์หรือคิดว่าอาจจะตั้งครรภ์ (สำหรับผู้หญิง) ณ เวลานี้		✓		
21. โรคไตหรือโรคที่เกี่ยวข้องด้านไต		✓		
22. อื่นๆ ถ้ามี (เช่น แพ้ยาประเภท ฯลฯ) ระบุ.....				

ประวัติการเจ็บป่วยของบุคคลในครอบครัว

ประวัติการเจ็บป่วย	มี	ไม่มี
1. โรคมะเร็งต่าง ๆ		✓
2. โรคหัวใจ		✓
3. โรคเบาหวานหรือความดันโลหิต	✓	
4. โรคหลอดเลือดในสมองอุดตัน,ตีบหรือแตก		✓
5. โรคทางพันธุกรรมอื่น ๆ ระบุ		
6. โรคอื่น ๆ ระบุ		

กรุณาลงเครื่องหมาย ✓ ในช่องด้านล่างนี้

*** ท่านยินยอมให้สามารถนำข้อมูลทั้งหมดในเอกสารฉบับนี้เพื่อใช้ในงานด้านสถิติหรืองานด้านวิจัยหรือไม่ ****

(ข้อมูลในส่วนนี้มีความจำเป็น กรุณาลงข้อมูลตามความเห็นชอบของท่าน)

☑ ยินยอม

☐ ไม่ยินยอม

ระดับความดันโลหิต (เริ่มต้น) / น้ำหนักและส่วนสูง

BP 122/96 mmHg. , น้ำหนัก ..... กก.

HR 106 ครั้ง/นาที , ส่วนสูง ..... ซม.

(เฉพาะเจ้าหน้าที่เท่านั้น)

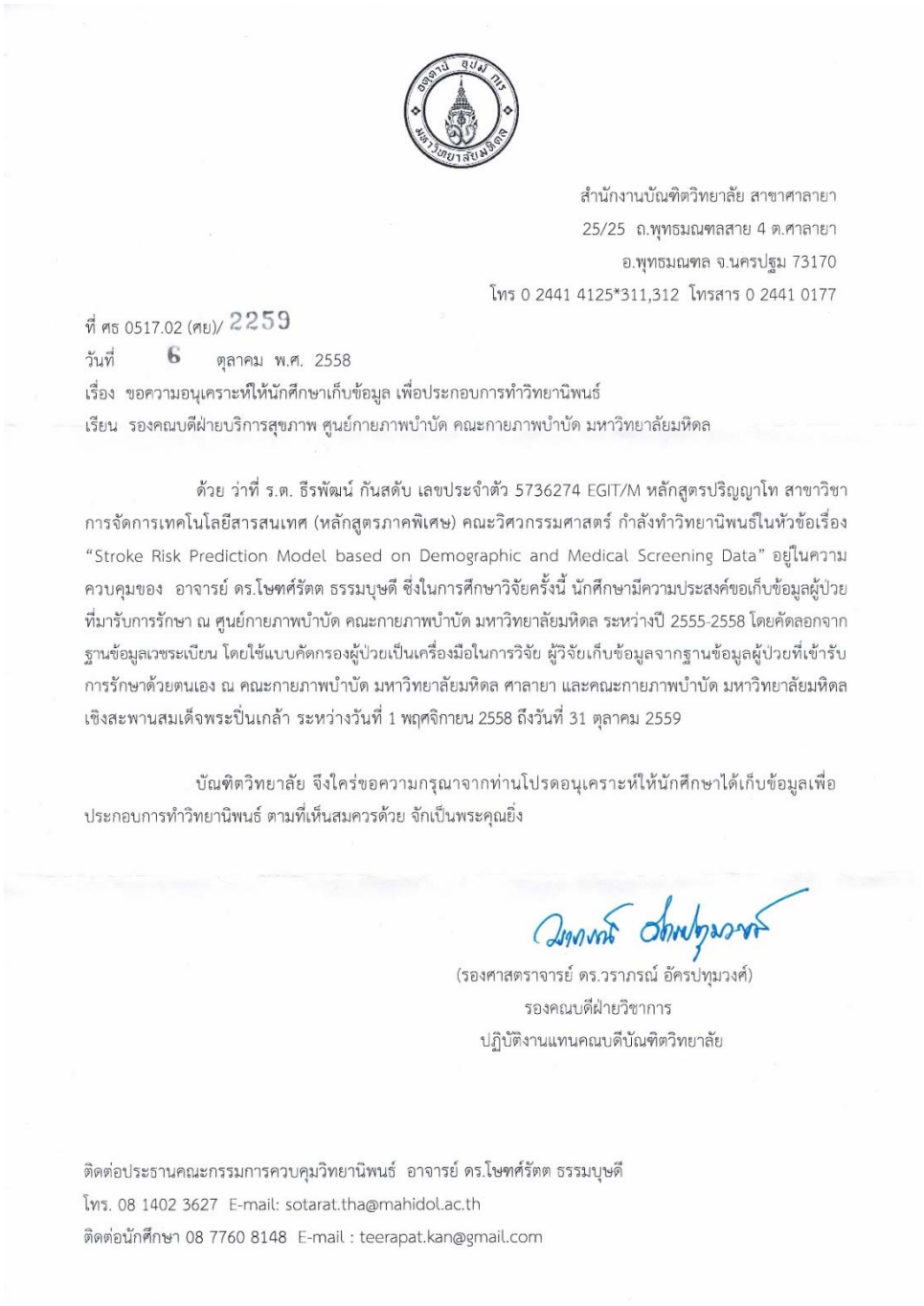
ลงชื่อ [redacted] (ผู้ป่วยหรือแทนผู้ป่วย)

วันเดือนปี ..... / ..... /..... (ที่ขอมีเวชระเบียน)

Figure A.4 Medical Screening Form since 2015.

APPENDIX B

APPROVAL FROM INSTITUTIONAL REVIEW BOARD



สำนักงานบัณฑิตวิทยาลัย สาขาศาลายา
25/25 ถ.พุทธมณฑลสาย 4 ต.ศาลายา
อ.พุทธมณฑล จ.นครปฐม 73170
โทร 0 2441 4125*311,312 โทรสาร 0 2441 0177

ที่ ศธ 0517.02 (ศย)/ 2259
วันที่ 6 ตุลาคม พ.ศ. 2558
เรื่อง ขอความอนุเคราะห์ให้นักศึกษาเก็บข้อมูล เพื่อประกอบการทำวิทยานิพนธ์
เรียน รองคณบดีฝ่ายบริการสุขภาพ ศูนย์กายภาพบำบัด คณะกายภาพบำบัด มหาวิทยาลัยมหิดล

ด้วย ว่าที่ ร.ต. ธีรพัฒน์ กันสดับ เลขประจำตัว 5736274 EGIT/M หลักสูตรปริญญาโท สาขาวิชาการจัดการเทคโนโลยีสารสนเทศ (หลักสูตรภาคพิเศษ) คณะวิศวกรรมศาสตร์ กำลังทำวิทยานิพนธ์ในหัวข้อเรื่อง "Stroke Risk Prediction Model based on Demographic and Medical Screening Data" อยู่ในความควบคุมของ อาจารย์ ดร.โชทศ์รัตต ธรรมบุษดี ซึ่งในการศึกษาวิจัยครั้งนี้ นักศึกษามีความประสงค์ขอเก็บข้อมูลผู้ป่วยที่มารับการรักษา ณ ศูนย์กายภาพบำบัด คณะกายภาพบำบัด มหาวิทยาลัยมหิดล ระหว่างปี 2555-2558 โดยคัดลอกจากฐานข้อมูลเวชระเบียน โดยใช้แบบคัดกรองผู้ป่วยเป็นเครื่องมือในการวิจัย ผู้วิจัยเก็บข้อมูลจากฐานข้อมูลผู้ป่วยที่เข้ารับการรักษาด้วยตนเอง ณ คณะกายภาพบำบัด มหาวิทยาลัยมหิดล ศาลายา และคณะกายภาพบำบัด มหาวิทยาลัยมหิดล เชิงสะพานสมเด็จพระปิ่นเกล้า ระหว่างวันที่ 1 พฤศจิกายน 2558 ถึงวันที่ 31 ตุลาคม 2559

บัณฑิตวิทยาลัย จึงใคร่ขอความกรุณาจากท่านโปรดอนุเคราะห์ให้นักศึกษาได้เก็บข้อมูลเพื่อประกอบการทำวิทยานิพนธ์ ตามที่เห็นสมควรด้วย จักเป็นพระคุณยิ่ง

(รองศาสตราจารย์ ดร.วราภรณ์ อัครปทุมวงศ์)
รองคณบดีฝ่ายวิชาการ
ปฏิบัติงานแทนคณบดีบัณฑิตวิทยาลัย

ติดต่อประธานคณะกรรมการควบคุมวิทยานิพนธ์ อาจารย์ ดร.โชทศ์รัตต ธรรมบุษดี
โทร. 08 1402 3627 E-mail: sotarat.tha@mahidol.ac.th
ติดต่อนักศึกษา 08 7760 8148 E-mail : teerapat.kan@gmail.com

Figure B.1 Data Request Form.

COA No. MU-CIRB 2015/127.2610

Mahidol University Central Institutional Review Board (MU-CIRB)

Certificate of Approval

Protocol No.: MU-CIRB : 2015/144.2509

Title of Project: Stroke Risk Prediction Model based on Demographic and Medical Screening Data

Approval includes:

1) Principle Investigator : Dr. Sotarat Thammaboosadee
   Affiliation : Faculty of Engineering, Mahidol University
   Research Site : Faculty of Physical Therapy, Mahidol University
2) Submission form version date 13 Oct 2015
3) Protocol version date 28 Sep 2015
4) Questionnaire version date 28 Sep 2015

MU-CIRB is in full compliance with International Guidelines for Human Research Protection such as Declaration of Helsinki, The Belmont Report, CIOMS Guidelines and the International Conference on Harmonization in Good Clinical Practice (ICH-GCP)

Date of Approval: *12 Nov 2015*
Date of Expiration: *11 Nov 2016*

Signature of Chairperson: .....
(Professor Dr. Rutja Phuphaibul)
MU-CIRB Chair

Signature of Institute Representative: .....
(Professor Dr. Sansanee Chaiyaroj)
Vice President for Research

* See *list of Co-Investigators at the back page*

Page 1 of 2

Figure B.2 Approval Form from IRB.

BIOGRAPHY

NAME	Acting Sub Lt. Teerapat Kansadub
DATE OF BIRTH	22 May 1990
PLACE OF BIRTH	Bangkok, Thailand
INSTITUTIONS ATTENDED	Rajamangala University, 2009-2012 Bachelor of Business Administration (Computer Information System) Mahidol University, 2015-2016 Master of Science (Information Technology Management)
HOME ADDRESS	27/130 M.6 Phuttamonthon Sai 4 Rd., Kratumrom, Samphran, Nakornpathom, Thailand, 73220 Tel. 087-760-8148 E-mail : teerapat.kan@gmail.com
PUBLICATION / PRESENTATION	Kansadub, T., Thammaboosadee, S., Kiattisin, S., and Jalayondeja, C. (2015), Stroke Risk Prediction Model based on Demographic Data, The Proceedings of the 2015 Biomedical Engineering International Conference (BMEiCON 2015), pp. 1-3. DOI: 10.1109/BMEiCON.2015.7399556.