Statistical and Machine Learning Models for Predicting of Hospital Cost on Diagnosis Related Groups (DRGs) in Chronic Disease in Southern Thailand

Wichayaporn Thongpeth

A Thesis Submitted in Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Research Methodology

Prince of Songkla University

2022

Copyright of Prince of Songkla University

Thesis Title	Statistical and Machine Learning Models for Predicting of Hospital Cost on Diagnosis Related Groups (DRGs) in Chronic Disease in Southern Thailand
Author	Mrs. Wichayaporn Thongpeth
Major Program	Research Methodology

Major Advisor:

Don McNeil
....
(Emeritus Prof. Dr. Don McNeil)

Co-advisors:

....
(Assoc. Prof. Dr. Apiradee Lim)

Examining Committee:

Chetta Ngamjarus
....Chairperson
(Assoc. Prof. Dr. Chetta Ngamjarus)

Don McNeil
....
(Emeritus Prof. Dr. Don McNeil)

....
(Assoc. Prof. Dr. Apiradee Lim)

....
(Asst. Prof. Dr. Phattrawan Tongkumchum)

....
(Asst. Prof. Dr. Salang Musikasuwan)

....
(Dr. Arinda Ma-a-lee)

The Graduate School, Prince of Songkla University, has approved this thesis as fulfillment of the requirements for the Doctor of Philosophy Degree in Research Methodology.

....

(Prof. Dr. Damrongsak Faroongsarng)
Dean of Graduate School

This is to certify that the work here submitted is the result of the candidate's own investigations. Due acknowledgement has been made of any assistance received.

Don McNeil
....Signature
(Emeritus Prof. Dr. Don McNeil)
Major Advisor

....Signature
(Mrs. Wichayaporn Thongpeth)
Candidate

I hereby certify that this work has not been accepted in substance for any degree, and is not being currently submitted in candidature for any degree.

....Signature
(Mrs. Wichayaporn Thongpeth)
Candidate

ชื่อวิทยานิพนธ์ ตัวแบบทางสถิติและการเรียนรู้ด้วยเครื่องสำหรับทำนายต้นทุนค่ารักษาพยาบาลตามระบบกลุ่มวินิจฉัยโรคร่วมในโรคเรื้อรังในภาคใต้ของประเทศไทย

ผู้เขียน นางวิชยาพร ทองเพชร

สาขาวิชา วิธีวิทยาการวิจัย

ปีการศึกษา 2564

บทคัดย่อ

วิทยานิพนธ์นี้เป็นการประยุกต์ใช้วิธีการทางสถิติในการสร้างตัวแบบทางสถิติสำหรับทำนายต้นทุนค่ารักษาพยาบาลตามระบบกลุมวินิจฉัยโรครวม (DRGs) ของโรคเรื้อรังในประเทศไทยและเปรียบเทียบประสิทธิภาพในการทำนายต้นทุนค่ารักษาพยาบาลตามระบบกลุมวินิจฉัยโรครวม (DRGs) ของโรคเรื้อรังระหว่างตัวแบบทางสถิติและตัวแบบการเรียนรู้ด้วยเครื่อง โดยแบ่งการศึกษาออกเป็นสองส่วน ดังนี้

ส่วนที่หนึ่งของการศึกษามีวัตถุประสงค์เพื่อศึกษาความสัมพันธ์ของปัจจัยที่มีผลต่อต้นทุนโรงพยาบาลตามระบบกลุ่มวินิจฉัยโรครวม (DRGs) ของโรคเรื้อรังและสร้างตัวแบบในการทำนายต้นทุนค่ารักษาพยาบาลตามระบบกลุ่มวินิจฉัยโรครวมโดยใช้ข้อมูลจากฐานข้อมูลผู้ป่วยในของโรงพยาบาลสุราษฎร์ธานีที่ใช้ในการเบิกจ่ายกับหลักประกันสุขภาพแห่งชาติ จำนวนที่เข้ารับการรักษารวมทั้งสิ้น 18,342 ครั้ง ตัวแปรที่ใช้ในการทำนายต้นทุนค่ารักษาพยาบาลตามระบบกลุ่มวินิจฉัยโรครวม คือ อายุ เพศ การวินิจฉัยโรคหลัก จำนวนการวินิจฉัยโรคแทรกซ้อน จำนวนหัตถการและการรักษา สถานภาพการจำหนายผู้ปวย จำนวนวันนอนในโรงพยาบาล ค่าใช้จ่ายในการรักษา ทำการวิเคราะห์ความสัมพันธ์ระหว่างปัจจัยทำนายและตัวแปรตามด้วยตัวแบบการถดถอยเชิงเส้น

ผลการศึกษาพบว่าปัจจัยที่มีผลต่อต้นทุนค่ารักษาพยาบาลตามระบบกลุ่มวินิจฉัยโรครวม มีความสัมพันธ์กับต้นทุนโรงพยาบาลในระบบกลุ่มวินิจฉัยโรคร่วมในโรคเรื้อรัง โดยมีค่า r^2 เท่ากับ 0.73 และปัจจัยที่มีความสัมพันธ์กับต้นทุนค่ารักษาพยาบาลตามระบบกลุ่มวินิจฉัยโรครวมในระดับที่สูง คือ จำนวนการทำหัตถการและการรักษา (r^2=0.54) และจำนวนวันนอนในโรงพยาบาล (r^2 = 0.43) โดยสรุป ปัจจัยหลักที่กำหนดค่ารักษาพยาบาลตามกลุ่มวินิจฉัยโรคร่วมในโรคเรื้อรัง คือ จำนวนการทำหัตถการและการรักษา และจำนวนวันนอนในโรงพยาบาล

ส่วนที่สองของการศึกษานี้ มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของแบบจำลองในทำนายต้นทุนค่ารักษาพยาบาลตามระบบกลุ่มวินิจฉัยโรครวมด้วยตัวแบบทางสถิติแบบเชิงเส้น (Linear Regression: LR) วิธีการถดถอยเชิงเส้นที่ปรับด้วยฟังก์ชันการลงโทษ (Penalized Linear Regression) ประกอบด้วย การถดถอยเชิงเส้นด้วยวิธีริดจ์ (Ridge Regression) การถดถอย

เชิงเส้นด้วยวิธีแลซโซ (Lasso Regression) วิธีการถดถอยอิลาสติคเน็ต (Elastic Net Regression) และตัวแบบการเรียนรู้ด้วยเครื่อง (Machine Learning: ML) ประกอบด้วยการเทคนิคซัพพอร์ตเวกเตอร์รีเกรสชัน (Support Vector Regression: SVR) โครงข่ายประสาทเทียม (Neural Network: NN) และป่าสุ่ม (Random Forest: RF) และ เอ็กซ์ทรีมกาเดียนบูทติ้ง (Extreme Gradient Boosting: XGBoost) ทำการแบ่งกลุ่มข้อมูลเป็นชุดข้อมูลเรียนรู้ และชุดข้อมูลทดสอบ ในสัดส่วน 70:30 และเพิ่มขนาดข้อมูลโดยวิธีบูตสแตรป (bootstrap) 2 เท่าและ 4 เท่า และวัดประสิทธิภาพการทำนายของแบบจำลองทั้งหมดด้วยค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (Root mean square error: RMSE) และสัมประสิทธิ์การกำหนด (Coefficient of determination: r^2)

ผลการศึกษาพบว่าการวิเคราะห์แบบวิธีป่าสุ่มให้ประสิทธิภาพของการทำนายดีที่สุดทั้งในข้อมูลที่ไม่ได้เพิ่มและเพิ่มขนาดตัวอย่าง โดยประสิทธิภาพการทำนายดีขึ้นเมื่อข้อมูลมีขนาดใหญ่ขึ้น ในขณะที่แบบจำลองทางสถิติ วิธีการถดถอยเชิงเส้นที่ปรับด้วยฟังก์ชันการลงโทษและเทคนิคซัพพอร์ตเวกเตอร์รีเกรสชันให้ประสิทธิภาพการทำนายใกล้เคียงกันสำหรับข้อมูลที่ไม่ได้มีการเพิ่มขนาดตัวอย่าง

โดยสรุปการถดถอยเชิงเส้นและการถดถอยเชิงเส้นที่ปรับด้วยฟังก์ชันการลงโทษ มีประสิทธิภาพในการทำนายใกล้เคียงกันและไม่เปลี่ยนแปลงสำหรับข้อมูลทั้งที่เพิ่มและไม่เพิ่มขนาด ส่วนแบบจำลองการเรียนรู้ด้วยเครื่องมีประสิทธิภาพดีขึ้นเมื่อขนาดข้อมูลใหญ่ขึ้น

Thesis Title	Statistical and Machine Learning Models for Predicting of Hospital Cost on Diagnosis Related Groups (DRGs) in Chronic Disease in Southern Thailand
Author	Mrs. Wichayaporn Thongpeth
Major Program	Research Methodology
Academic Year	2021

ABSTRACT

This dissertation applied statistical methods for predicting hospital cost on Diagnosis Related Groups (DRGs) for chronic disease in Southern Thailand. This study consists of two parts.

The first part of this dissertation aimed to analyze the determinants of costs for chronic disease patient visits in a major public hospital based on hospital claim data from Suratthani hospital in 2016. There was a total of 18,342 records of hospital visit costs. The determinant for predicting hospital cost included age and gender, principal and up to 12 diagnoses, up to 12 number of procedures, length of stays and discharge status. Linear regression was used to analyze associations between determinants and outcome. This study shows that the hospital cost determinants for chronic disease patients were the number of procedures (r^2=0.54) and length of hospital stay (r^2 = 0.43) with r^2 of 0.73. In conclusion, the main factors effected hospital costs for chronic disease are the number of procedures and length of hospital stay.

The objective of the second part of this dissertation was to compare linear regression, penalized linear: including lasso ridge and elastic net and machine learning models: including support vector regression (SVR), neural network (NN) random forest (RF), and Extreme Gradient Boosting (XGBoost) prediction performance of hospital visit cost from chronic disease in Thailand. The original data was divided into a training and testing set with 70:30 ratios and a double-sized dataset produced by the bootstrap technique. All models' predictive performance was measured with root mean square error (RMSE) and the Coefficient of determination (r^2).

The results revealed that the RF model had the best predictive performance of hospital visit cost for all dataset sizes in training and testing datasets with the lowest prediction errors. In contrast, linear regression had the most inadequate prediction performance and the highest prediction errors. RF, XGBoost, NN, and SVR models had better prediction performance for larger samples except for the linear regression model and penalized linear.

In conclusion, linear regression and penalized linear models had similar prediction performance for all sample sizes, whereas machine learning had better performance when the sample size increased.

ACKNOWLEDGEMENT

First and foremost, I would like to express my sincere gratitude to my supervisors, Emeritus Prof. Dr. Don McNeil and Assoc. Prof. Dr. Apiradee Lim for their helpful guidance, continuous support and invaluable assistance throughout the completion of the thesis, for their patience, motivation, enthusiasm, most importantly, and immense knowledge. Their guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better supervisor and mentor for my Ph.D. study. Furthermore, thanks to the staff in the Research Methodology at the Department of Mathematics and Computer Science, Faculty of Science and Technology, Prince of Songkla University, Pattani Campus consists of Asst. Prof. Dr. Phattawan Tongkumchum and Asst. Prof. Dr. Salang Musikasuwan for their encouragement and suggestion.

This thesis would never have been completed without the essential data from Suratthani Hospital. Furthermore, I would like to acknowledge financial support from the Graduate School, Prince of Songkla University, for Thailand's research, authorship, and publication.

Finally, I would like to express my heartfelt thanks to my parents for giving birth to me in the first place and supporting me with unfailing love, caring, encouragement throughout my life and for all of the benefits arising from this study. Similarly, I wish to thank my parents, mother, father, husband, and two daughters, who have never ended encouragement in all aspects and keep me hopeful.

Wichayaporn Thongpeth

TABLE OF CONTENTS

Page

TABLE OF CONTENTS (Continued)

Page

TABLE OF CONTENTS (Continued)

Page

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

LIST OF FIGURES (Continued)

Figure **Page**

LIST OF ABBREVIATIONS AND SYMBOLS

DRGs	=	Diagnosis-Related Groups
NHSO	=	National Health Security Office
WHO	=	World Health Organization
LOS	=	Length of hospital stay
nProc	=	Number of procedures
nDiag	=	Number of diagnoses
HIV	=	Human Immunodeficiency Virus
ICD-9	=	International Classification of Diseases 9
ICD-10	=	International Classification of Diseases 10
LR	=	Linear regression
Penalized LR	=	Penalized linear regression
ML	=	Machine Learning
RF	=	Random Forest
SVR	=	Support Vector Regression
NN	=	Neural Network
XGBoost	=	Extreme Gradient Boosting
RMSE	=	Root Mean Square Errors
r^2	=	The coefficient of determination

CHAPTER 1

INTRODUCTION

1.1 Overview of the thesis

This PhD dissertation focuses on using statistical methods to create an appropriate model for predicting hospital cost on Diagnosis Related Groups (DRGs) in chronic disease in Southern Thailand and comparing statistical models with machine learning algorithms to predict hospital cost. The Thailand National Health Security Office (NHSO) collects data on all hospital visits by patients with chronic illnesses, including age, gender, discharge status, primary and 12 secondary diagnoses, up to 12 treatment procedures, and total visit expenses. All subjects involved in this study contained data from 2016 on chronic patients of DRGs, and data were obtained from a Suratthani hospital. This thesis is divided into four chapters:

Chapter 1 is a thesis introduction that includes a rationale, objectives, scope of the study and literature review.

Chapter 2 describes the methodology, including study design, data sources, data management, statistical analysis, and conceptual framework.

Chapter 3 illustrates and reports on statistical methods for determining hospital costs of DRGs in chronic disease in Thailand.

Chapter 4 discusses the results and statistical analysis methods used and applications evident from this study.

1.2 Introduction

Providing healthcare services is an essential part of the global economy. According to the World Bank, global health expenditures accounted for 10% of the global gross domestic product in 2016 (WHO, 2019). It is critical to understand how production levels and other variables affect hospital costs. Cost analysis enables department heads, hospital administrators, and policymakers to assess their institutions' ability to meet these public needs (Chilingerian *et al*., 2008). Apart from that, the administrators will utilize those limited resources efficiently and effectively. The accuracy of the cost information is essential to correct decision-making (Mihailovic *et al.*, 2016). Government hospitals in developing and developed countries should be managed for the greater benefit of the community. Higher hospital cost relates to higher utilization of hospital resources and the severity of diseases (Pritchard *et al*., 2016; Yang *et al.*, 2018).

The DRGs are a classification system for patients that generally cover all inpatient stay costs from admission to discharge. Hospitals in most developed countries have been using DRGs as a tool for assessing reimbursement for over 30 years (Schreyögg *et al*., 2006; Chilingerian, 2008; Scheller-Kreinsen *et al.,* 2009; Mihailovic *et al*., 2016; Briestensky *et al.,* 2021).

Mathauer and Wittenbecher (2013) advised that the methods to estimate hospital costs in low- and middle-income nations with limited resources may differ. However, the significant variables were influencing hospital prices in wealthy countries. DRGs are used to determine hospital reimbursement for acute inpatient care and chronic inpatient treatment. According to Ding *et al.* (2017), chronic diseases cost on average three times as much as acute diseases for inpatients. The prevalence of chronic illnesses is growing, resulting in a significant increase in health care expenditures. It demonstrates the critical nature of disease prediction, which is essential for both government and insurance corporations when developing health care budgets and insurance programs (Sav *et al.,* 2015; Bernell and Howard., 2016; Pritchard *et al.,* 2016; Lin *et al.,* 2018; Toxvaerd *et al.,* 2019; Holman, 2020).

Health care expenses are increasing in Thailand, with one of the key reasons being the frequency of chronic diseases. The NHSO is the major purchaser of health care in the nation through the tax-financed Universal Coverage Scheme (UCS), which distributes pooled monies to health providers. Many hospitals face financial difficulties because they incur higher medical care costs than they receive from the NHSO. Even though the NHSO has been using DRGs for more than a decade, several issues remain unresolved (Pongpirul *et al.,* 2011; Sakunphanit, 2015). This study aimed to determine factors used in the DRGs system related to hospital costs and compare statistical with machine learning models for predicting hospital costs. Typically, data on health care costs are positively skewed. Machine learning (ML) has made significant strides over the last three decades, and these models have recently been applied to a variety of healthcare datasets. Only a few comparisons of the prediction performance of linear regression (LR) and ML models on highly positively skewed data, such as health care costs, have been conducted in Thailand using different training sample sizes. However, studies that compare the predictive performance of LR and ML models with varying sample sizes for substantially favorably skewed data such as healthcare expenses are limited (Sushmita *et al.,* 2015; Panay *et al.,* 2019; Kan *et al.,* 2019; Hanafy and

Mahmoud, 2021). Therefore, this study aimed to apply the statistical model for predicting hospital cost and compare the performance of the statistical models and machine learning models.

1.3 Objectives

1. To apply the statistical models to predict hospital costs from chronic disease patient visits in Southern Thailand

2. To compare the predictive performance of standard LR, penalized LR including lasso, ridge, and elastic net, and four types of ML models, namely support vector regression (SVR), neural network (NN), random forest (RF) and Extreme Gradient Boosting (XGBoost) models

1.4 Scope of the study

This study analyzed secondary data from the Surat Thani tertiary hospital database in southern Thailand, which included claims for health care costs per capita from the Thailand National Health Security Office. The natural logarithm of hospital costs was taken to reduce skewness by adding 1 to avoid zero cost. The predictors of hospital cost included patient's age, gender, treatment outcome, number of diagnoses and secondary diagnoses ranging from 0 to 12, number of procedures ranging from 0 to 12. The prediction performance standard LR and penalized LR and ML models were compared. We employed the root mean square error (RMSE) and coefficient of determination (r^2) to determine the optimal model.

1.5 Literature Review

Several publications relevant to this study were reviewed, including the statistical methods used and their findings. In addition, DRGs, chronic disease, and the statistical techniques used for these studies have been reviewed.

1.5.1 The Diagnosis-Related Groups (DRGs)

The definition and background of DRGs

In most nations, health care expenses are governed mainly by the 1983-instituted DRGs system (Mihailovic *et al.,* 2016). DRGs are a classification system for patients based on standardized prospective payments to hospitals that typically cover all costs associated with an inpatient stay from admission to discharge. Patients' primary and secondary diagnosis, surgical procedures, comorbidities and complications, age, gender, and treatment outcome determine their DRGs classification (Scheller-Kreinsen *et al.,* 2009). The DRGs system is used to control costs, improve the efficiency, transparency, and equity of health financing, and assist hospitals in their administration (Busse *et al.,* 2011). Over the last 30 years, hospitals in most industrialized nations have used DRGs to determine payment. Under this arrangement, the paying party of medical insurance does not pay for inpatients' actual expenses but is based on DRGs (Scheller-Kreinsen *et al.,* 2013; Choi *et al.,* 2019). Individual patient utilization of hospital outputs is dependent on both the patient's condition and the treatment procedures used. DRGs and hospital costs are frequently used parameters to indicate utilization of health resources, health care costs, and disease severity (Lee *et al.,* 2004; Gartner *et al.,* 2015; Liu *et al.,* 2018).

Cashin *et al.* (2005) argued that the approaches for estimating hospital costs in low- and middle-income resource-poor countries are distinct, even though the primary determinants affecting hospital costs are the same as in developed countries. Thus, it is necessary to evaluate DRGs-based payments in these nations to improve health care efficiency, equality, and quality (Kankeu *et al.,* 2013). DRGs were first used to determine hospital payment for acute inpatient treatment and determine costs for chronic inpatient care globally (Hendriks *et al.*, 2014; Chapel *et al.*, 2017). Yuan *et al.* (2019) investigated the impact of the DRGs payment reform on the global budget in Zhongshan, China. They suggested DRGs positively affected Acute Myocardial Infraction (AMI) patients' cost containment, but the effects on resource utilization were negative.

Thailand's NHSO is the primary purchaser of health care nationwide, covering 76.0 per cent of the population via a tax-financed UCS that distributes pooled money to health providers (Tangcharoensathien *et al.,* 2018). The NHSO has been using the DRGs system for over a decade, although many problems remain unresolved. Additionally, many hospitals are presently experiencing financial difficulties since their medical care expenses exceed the compensation received from the NHSO. One potential explanation for Thailand's numerous financial problems is that the existing DRGs do not accurately reflect the actual cost of medical treatment. Identifying the significant factors affecting hospital costs assists policymakers inequitable allocation and efficient reimbursement of funds to health providers. However, no such study of costs has been conducted recently, and thus the system lacks sufficient data for informed analysis of the current situation. As a result, this research sought to determine the factors influencing the cost of visits to a large public hospital by patients with chronic illnesses.

1.5.2 Chronic disease

Globally, health care costs continue to rise. One of the primary contributors to the rise in chronic illness is because it is more severe and frequently incurable through vaccinations or other medicines (Dans *et al.,* 2011). In the United States (US), chronic diseases affect approximately half of the population. In Europe, the rising number of chronic illnesses accounts for 75% of total healthcare expenditures, respectively (Kerr *et al.,* 2007). The top 5% of patients are responsible for half of all health care spending, while the top 1% of spenders are responsible for almost 27% of costs (Glynn *et al.,* 2011; Toxvaerd *et al.,* 2019; Holman, 2020).

Chronic diseases could cost up to 7% of the gross domestic product of any country due to the detrimental effects on economic activity and the increased expenditure on public health and social welfare. By 2030, it is predicted that these diseases will cost China $7.7 trillion, Japan $3.5 trillion, and South Korea $1 trillion (Miranda *et al.,* 2008; Thorpe and Philyaw, 2012; Pritchard *et al.,* 2016; Bloom, 2017; Lin *et al.,* 2018).

Cardiovascular illnesses, cancer, chronic respiratory diseases, and diabetes are the top four chronic diseases that cause the most fatalities worldwide, particularly in low- and middle-income nations (WHO, 2019). Patients with chronic conditions face health and physical limitations and the financial burden of disease care. Chronic disease complications, such as diabetes, are expensive and rarely treatable (Collins *et al.,* 2009; Kankeu *et al.,* 2013; WHO, 2019). Chronic illnesses are the most significant cause of mortality globally; however, they can be partly prevented by interventions (Glasgow *et al.,* 2001; Meetoo, 2008). The need for models capable of predicting healthcare expenses is critical (Hansen, 2016). High-cost patients are the most expensive patients globally, and high levels of chronic illness mainly explain their high utilization, a study has found. Preventable spending on health should be maximally 10% of annual income, according to Wammes *et al.* (2018).

Bredenkamp *et al.* (2020) discovered that DRGs payments are frequently used for daycare and surgery services. Exclusions may include costly medications, sophisticated therapies, transplants, emergency care, psychiatry, rehabilitation, long-term nursing care, tuberculosis, and HIV/AIDS cases. Wu *et al*. (2020) examined the composition and factors of inpatient hospital costs for colorectal surgery. They conducted a study in Beijing on the usage of DRGs. They discovered that age, gender, length of stay (LOS), diagnosis, treatment, and clinical procedures significantly impacted the inpatient cost of colon cancer patients in China.

1.5.3 Determinants of hospital cost from DRGs system

Hospital cost generally means the financial liabilities hospitals incur by providing care to patients. Usually, the determinants of hospital costs are identified from the factors used in the DRGs system (Malehi *et al.,* 2015). Age, gender, principal diagnosis, secondary diagnosis, a surgical procedure performed, comorbidities and complications, and discharge status are used to assess patients in the DRGs system in order to determine hospital costs (Hansen, 2016; Bramkamp *et al*., 2007; Wu *et al.,* 2020). Several studies evaluated the appropriate use of these factors to determine the hospital cost (Evans *et al.,* 1995; Silber *et al.*, 1999; Penberthy *et al.,*1999; Warren *et al.*, 2008;

Chaikledkaew *et al.,* 2008; Hansen *et al.,* 2016; Bramkamp *et al.,* 2007; Liu *et al*., 2018; Wu *et al*., 2020).

Chaikledkaew *et al.* (2008) investigated the factors influencing healthcare expenses and hospitalizations in diabetes patients treated in Thai public hospitals. They discovered that increasing healthcare costs were significantly associated with patients' gender and age. Increases in all significant hospital costs for chronic diseases were associated with increasing age. The rising medical expenditures among older patients are more likely to have comorbidities requiring particular medication and more prolonged treatment (Peltola and Quentin, 2013; Angstman *et al*., 2016; Wu *et al*., 2020; Xu *et al.,* 2020). Several studies concluded that male patients had higher hospital costs (Krop *et al*., 1998; Aljunid and Jadoo, 2008).

In contrast, Owens (2008) studied gender differences in health care expenditures, resource utilization, and quality of care and revealed that women had substantially higher medical care costs than men. The difference in hospital cost is probably due to the type of disease. Female patients, especially 45–64 years of age, can absorb more hospital resources. They have gender-specific conditions, for instance, menopausal symptoms and prenatal conditions. Wu *et al.* (2020) examined the factors influencing hospital costs for colorectal cancer patients at a Beijing hospital. They discovered that male patients had lower inpatient expenditures than female patients and that age had a beneficial effect on inpatient costs.

Apart from patients' demographic factors, LOS and clinical factors play an essential role in determining hospital costs among chronic patients. The LOS has been identified as the primary predictor of hospital costs (Philbin *et al.,* 2001; Wu *et al.,* 2020). The more extended LOS higher hospital costs were reported by several studies (Evans *et al*., 1995; Krop *et al*., 1998; Philbin *et al*., 2001; Slabaugh *et al*., 2015; Nelson-Williams *et al*., 2016; Bramkamp *et al*., 2007; Aljunid and Jadoo, 2018; Kuo *et al.,* 2018; Liu *et al.,* 2018; Wu *et al.,* 2020). This is due to the resources used in the hospital when the patients spend a longer time. The longer a patient is in a hospital, the more hospital resources are spent. Indeed, prior research has revealed hospital strategies for accelerating or shortening patient LOS, legally or illegally, to contain inpatient expenditures, particularly under a DRGs payment system (Perelman and Closon, 2007; Hamada *et al.,* 2012).

Kuo *et al.* (2018) investigated models used to predict medical costs associated with spinal fusion surgery in Taiwan and discovered that LOS plays a role in determining medical expenditures. Nelson-Williams *et al.* (2016) examined the factors that contribute to hospitalisation costs for patients undergoing hepatopancreatic biliary surgery. The research indicated that more significant hospital costs mainly were associated with a longer LOS. Wu et al. (2020) used a decision tree model to assess the effect of LOS and other variables on colorectal cancer inpatient medical expenses. Their study established that the LOS and patient characteristics are significant predictors of medical expenditures.

The studies conducted in the United States and Taiwan reported that the comorbidities among diabetic patients were significantly associated with direct medical expenditures (Krop *et al.,* 1998; Guo *et al.,* 1998; Krop *et al.,* 1999; Bhattacharyya and Else., 1999; Brown *et al.,* 1999). Kim *et al*. (2004) evaluated trends in hospital utilization and expenses for HIV/AIDS patients in South Carolina from 1994 to 1996. When

HIV/AIDS is the primary admitting diagnosis, hospitalization expenditures are greater. Increased sickness severity (number of diagnoses) results in increased overall hospital charges and days. Increases in high hospital costs for chronic diseases were associated with increasing comorbidity and complication (Gordon *et al.*, 2012; Aljunid and Jadoo, 2017).

Treatment procedures were one of the most significant factors on hospital cost among chronic disease patients. The higher number of procedures causes higher medical costs, as reported by Silber *et al.* (1999), Philbin *et al.* (2001), Warren (2008), Benoit and Cohen (2001) and Wu *et al.* (2020). Aljunid and Jadoo (2017) investigated factors affecting healthcare costs and hospitalizations in Malaysian public hospitals' total inpatient pharmacy. They discovered that specific surgical treatments and serious complications cost more than a medical case. Wu *et al.* (2020) demonstrated that differences in therapy and LOS were significant determinants of inpatient medical spending in patients with colorectal cancer.

1.5.4 Hospital cost prediction performance from different models

The most common statistical model used for predicting hospital costs is the linear regression model. However, hospital costs are usually skewed positively (Dodd *et al.*, 2006; Gertman and Lowenstein, 1984). Therefore, the cost transformation before performing modelling analysis is needed (Ai and Norton, 2000; Duan *et al.*, 1983; Manning and Mullahy, 2001; Veazie *et al*., 2003; Gregori *et al*., 2011; Franzco *et al*., 2014). Currently, ML is gaining popularity because of better predictive performance with a larger sample size (Gilleskie and Mroz, 2004; Conigliani and Tancredi, 2009; Basu *et al.*, 2006; Hill and Miller, 2009; Mihaylova *et al*., 2011; Dureh and

Tongkumchum, 2019; Lim *et al.*, 2020). The volume of data from several hospitals is increasing and is considered big data. Therefore, ML has become an alternative method to predict hospital costs. Comparing the prediction performance between the traditional statistical model and ML had been performed by many studies (Kulkarni *et al.*, 2020; Austin *et al*., 2003; Bertsimas *et al*., 2008; Ding *et al*., 2017; Kuo *et al.*, 2018; Seligman *et al*., 2018; Kan *et al*., 2019; Patil *et al*., 2020). Traditional statistical models require assumptions, whereas ML models do not require assumptions, and the relationship between outcome and determinants need not be linear (Boulesteix and Schmid, 2014; Bzdok *et al*., 2018; Kourou *et al*., 2015). Furthermore, these ML models perform better when applied to a larger dataset (Povak *et al.*, 2014; Dureh and Tongkumchum, 2019; Sweety *et al.*, 2019; Lim *et al*., 2020; Rajula *et al.*, 2020). However, ML models may have the problem of overfitting.

Models such as Poisson regression, negative binomial, proportional hazards and gamma regression have been applied to predict medical cost and surgical treatment. Based on various model assessments, these models accurately predicted the cost of treatment under varying assumptions (Austin *et al.,* 2003). Apart from traditional statistical models, the number of studies using ML is increasing with the advancement of technology in keeping large datasets. For example, Kulkarni *et al*. (2020) used various ML algorithms to predict inpatient hospital charges. The ML algorithms included RF, stochastic gradient descent (SGD) regression, K-closest neighbour regressor, XGBoost regressor and gradient boosting regressor. Findings from this study indicated a significant positive correlation between hospital LOS and total cost. Among these algorithms, the RF achieved the highest predictive accuracy with an r^2 of 0.7753

Lee *et al.* (2004) have used ANN and classification and regression tree (CART) to predict hospital charges of colorectal cancer treatment. Based on their results, ANN models showed better accuracy in the linear correlation coefficient. Muremyi *et al.* (2019) predicted the out-of-pocket medical expenditures in Rwanda using four ML approaches: RF, decision tree (DT), gradient boosting machine (GBM), and regression tree models. They found that GBM has prediction efficiency and accuracy higher than other ML algorithms with r^2 0.853 and adjusted r^2 0.853. Sushmita *et al.* (2015) used regression trees, M5 model tree and random forest to predict healthcare costs of individual patients. M5 model accurately predicted costs within less than $125 for 75% of the population compared to prior techniques. Sushmita *et al.* (2016) employed ML algorithms to predict 30-day risk and expense based on admission data from a large hospital chain in the Northwestern U.S. They analyzed LOS, admission acuity level (A), comorbid conditions (C), and utilization of emergency departments (E) using LR and RF regression. The results indicated that RF had better prediction performance in a larger sample size.

Duncan *et al.* (2016) compared generalized linear models (GLMs), multivariate adaptive regression splines, RF, DT and boosted trees. Despite the short sample size and non-normal distribution, performance analysis revealed that advanced supervised ML outperform traditional regression models. Findings from research conducted by Yang *et al.* (2018) indicated that recurrent neural networks perform better at predicting the medical expenditure of high-cost, high-need patients and LR, Lasso and GBM. Kuo *et al.* (2018) discovered that the RF model was the most accurate in predicting the medical expenditures related to spinal fusion in Taiwan DRGs in terms of profit or loss.

Lakshmanarao *et al.* (2020) proposed a machine learning model for predicting medical costs. They observed that age, body mass index (BMI) are features that decide the dependent variable. Out of all experiments, RF has given better results than other methods.

From several previous studies, using ML overcame traditional statistical models in the larger sample size. However, the conclusion of the best model for prediction performance among ML models is sparse.

CHAPTER 2

METHODOLOGY

This chapter describes the overall research methodology used to predict hospital costs for DRGs in Southern Thailand. The data sources, management processes, and analysis are described in detail. Following that, the study population, sample, variables, and statistical analysis are described in two distinct sections (analyze the determinants and compare LR, penalized LR and ML models prediction performance of hospital costs for chronic-disease patient visits).

2.1 Study design

A retrospective data analysis was performed to predict hospital costs on DRGs in chronic disease in Southern Thailand.

2.2 Data sources

The Thailand NHSO collected data on all hospital visits by patients with chronic diseases, including admission and discharge dates, age, gender, discharge status, primary and 12 secondary diagnoses, up to 12 treatment procedures, and total visit costs. The study analyzed 18,506 hospital visits to Surat Thani regional hospital in 2016 by patients with chronic illnesses.

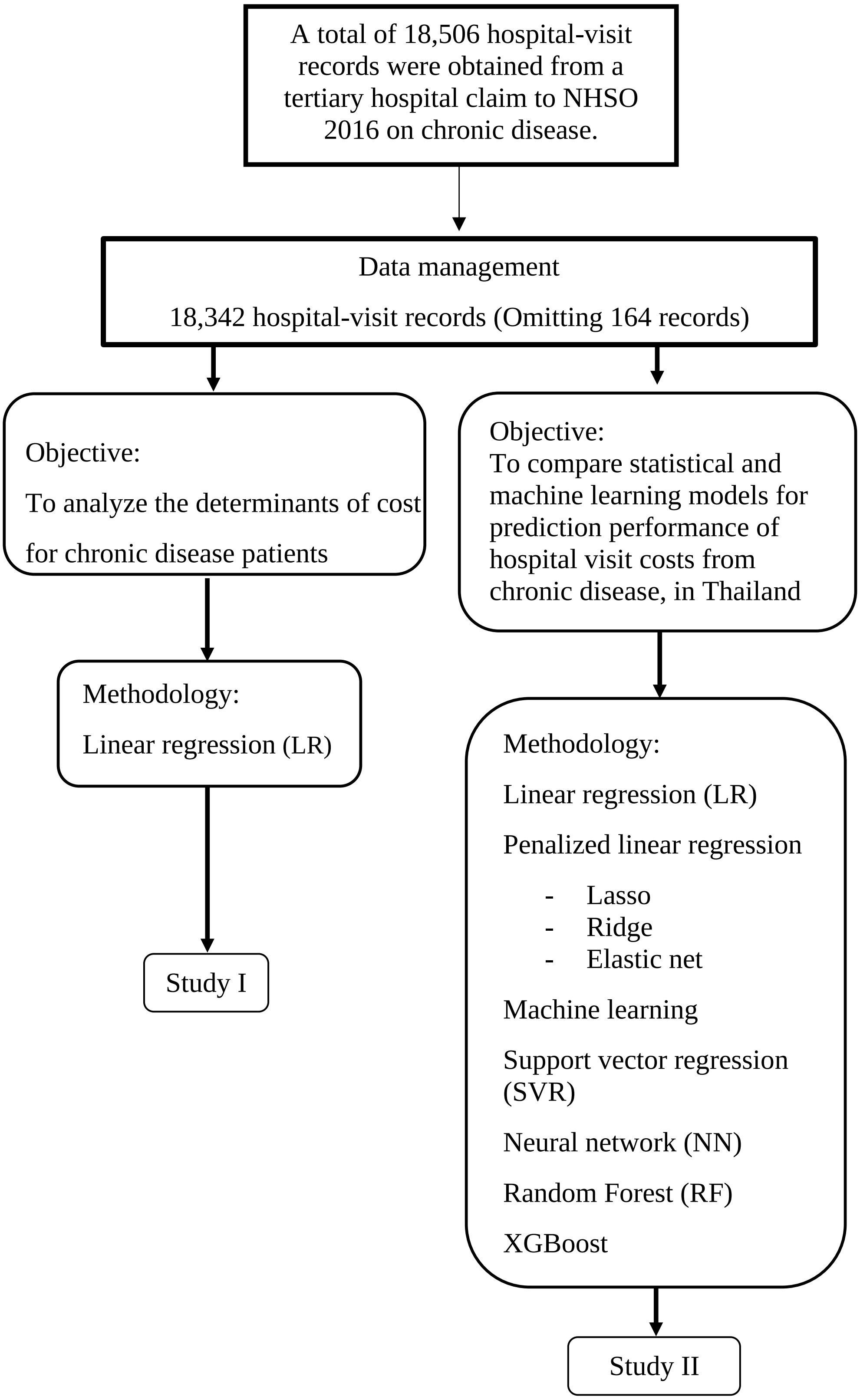

A total of 18,506 hospital-visit records were obtained from a tertiary hospital claim to NHSO 2016 on chronic disease.
Data management
18,342 hospital-visit records (Omitting 164 records)
Objective:
To analyze the determinants of cost for chronic disease patients
Objective:
To compare statistical and machine learning models for prediction performance of hospital visit costs from chronic disease, in Thailand
Methodology:
Linear regression (LR)
Methodology:
Linear regression (LR)
Penalized linear regression
- Lasso
- Ridge
- Elastic net
Machine learning
Support vector regression (SVR)
Neural network (NN)
Random Forest (RF)
XGBoost
Study I
Study II

Figure 2.1 Workflow of this study

2.3 Data management

In study I, Thailand's NHSO collects data on the costs of all hospital visits by patients with chronic conditions, including admission and discharge dates, age, gender, discharge status, primary diagnoses (ICD 10 group), complications or comorbidity, treatment procedures (ICD-9-CM) and total visit cost. The study collected 18,506 hospital visits to Surat Thani regional hospital in 2016 by individuals with chronic conditions. Data cleaning was undertaken to identify and remove errors and prevent duplication of records. As a result of the initial descriptive analysis of the complete sample using normal quantile-quantile plots plot of hospital cost as shown in Figure 3.1, all patient visits with abnormally medical expenditures less than 800 Baht (160 records) and more than 7 million Baht per day were eliminated, leaving 18,342 qualifying records for analysis. A log-linear model for estimating costs based on seven variables fitted the data were performed. For patients with chronic diseases, each visit's possible lowest healthcare costs are 800 Baht.

Males and females were classified separately. The age distribution was divided into ten groups of ten years each: 0-9 years, 10-19 years, 20-29 years, 30-39 years, 40-49 years, 50-59 years, 60-69 years, 70-79 years, 80-89 years, and 90 years and older. Gender and age categories were combined to create a new gender-age group variable with 20 categories. These two variables were then combined to reduce the interaction between gender and age. The length of stay in the hospital (LOS) was divided into 12 categories: 0, 1, 2, 3, 4, 5, 6, 7-8, 9-11, 12-15, 16-24 and 25 or more days. The principal diagnosis (ICD 10) was divided into 18 categories: tuberculosis, sepsis, HIV and other infectious illnesses, liver cancer, lung cancer, other digestive diseases, other cancers, endocrine disorders, muscle and neurological system disorders, ischemic heart disease, stroke, and other cardiovascular diseases, respiratory and digestive diseases, and genitourinary disorders, and ill-defined diseases. The following six discharge categories were used: denied, exited, escaped, other, death with an autopsy, and death without an autopsy. The number of diagnoses varied between one and thirteen, while the number of operations varied between zero and twelve.

In study II, the Suratthani tertiary hospital cost data excluded the cost less than 800 Baht and more than 7 million Baht, leaving 18,432 hospital visits were used for comparing prediction performance between standard statistical model, 3 penalized linear models (lasso, ridge and elastic net) and 4 ML models (SVR, NN, RF and XGBoost). The variables in this dataset include the patient's age, gender, admission date, discharge date, discharge status, the patient's number of primary and secondary diagnoses, which ranges from 0 to 12, the patient's number of treatment procedures, which also ranges from 0 to 12, and the total visit cost. Gender-age group, LOS, disease category, discharge status, number of diagnoses (nDiag), and number of procedures (nProc) are all considered determinants. Cleansing, integrating, and transforming data were all steps in the data preparation process.

2.4 Path diagrams and variables

Study I, hospital cost in Baht per patient visit, was taken as a natural logarithm divided by 100 and added one to prevent the problem of transformation from zero cost and was used as the outcome. The following path diagrams show the association between determinants and outcomes for both parts. Gender, age group (LOS), ICD-10 group, nDiag, nProc and discharge status are all considered determinants.

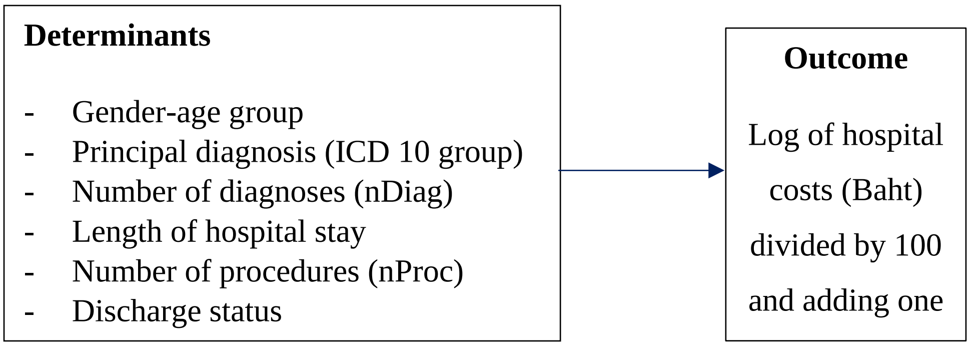

Determinants
- Gender-age group
- Principal diagnosis (ICD 10 group)
- Number of diagnoses (nDiag)
- Length of hospital stay
- Number of procedures (nProc)
- Discharge status
Outcome
Log of hospital costs (Baht) divided by 100 and adding one

Figure 2.2 Path diagram for Study I and II

Study II, the outcome is the natural logarithm of hospital cost divided by 100 and adding one. The determinants are the same as the determinants used in study I.

There were three different data forms used in this study. The first form was original data from hospital cost with a total sample size of 18,342, which was used for creating and evaluating the models and named as original data. The second form was original data separated into two sets randomly: training and testing sets with 70:30 ratios and named as split data. The last format was the data with two- and four-fold size by applying the bootstrap method. The original data were resampled. We separated these data into training and testing sets with the same ratio as the second form and named bootstrap data. The performance of the standard linear model penalized linear models and ML were compared using 3 sizes of datasets: original, data increased the size by two and four-fold. Each data size was split into two sets: training and testing set with the ratio of 70:30.

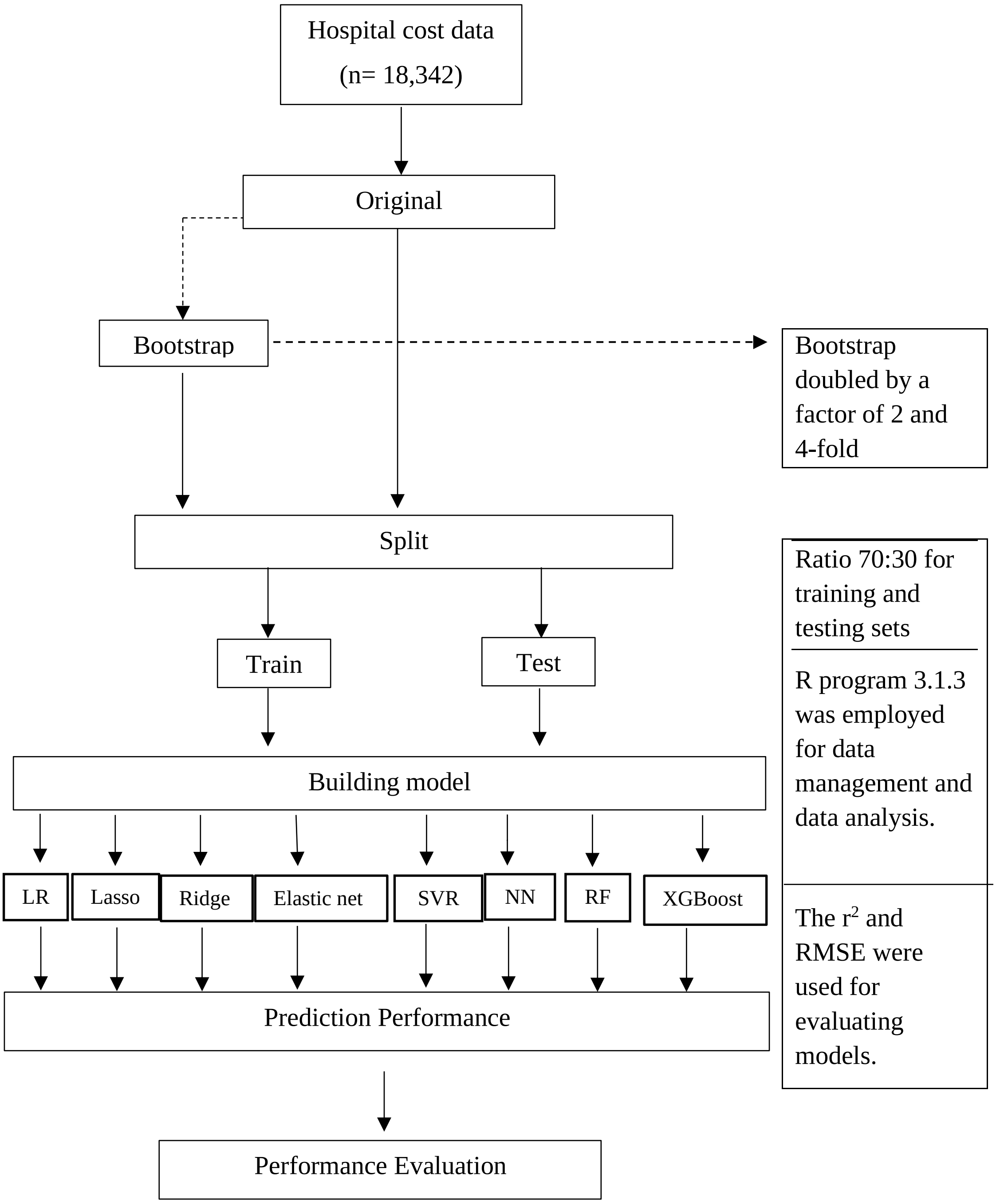

Hospital cost data
(n= 18,342)
Original
Bootstrap
Bootstrap doubled by a factor of 2 and 4-fold
Split
Ratio 70:30 for training and testing sets
R program 3.1.3 was employed for data management and data analysis.
Train
Test
Building model
LR
Lasso
Ridge
Elastic net
SVR
NN
RF
XGBoost
The r2 and RMSE were used for evaluating models.
Prediction Performance
Performance Evaluation

Figure 2.3 Conceptual framework of study II

2.5 Descriptive analysis

Study I

Descriptive analysis was performed for calculating counts and percentages for all variables.

Study II

For categorical data, descriptive analysis was used to calculate counts and percentages. The median and range of continuous variables were determined (hospital cost, number of procedures and length of hospital stay).

2.6 Predictive Models

2.6.1 Linear regression

Study I

Normal quantile-quantile plots were created to illustrate non-transformed and transformed hospital cost distribution. The relationship between cost and determinants were examined using multiple linear regression. The model's coefficients and standard errors were substituted in the linear equation to convert to be Baht. The 95 percent confidence interval (CI) graphs were created to demonstrate the multivariate analysis's results. Only significant components were included in the final model. The R software, version 3.1.3, was used to produce all statistical analyses and graphics (R Core Team, 2020).

Study II

For this study, the predictors of all models are gender-age group, LOS, ICD 10 group, discharge status, nDiag and nProc, and the outcome is hospital cost. The data were randomly separated into training and test sets with the ratio of 70:30.

The linear regression model (LR) is a well-established statistical technique for describing the relationship between a continuous outcome and categorical or continuous variables. A parametric model assumes a linear relationship between the outcome and the determinants. The errors are expected to be regularly distributed and constant variance. The model is as follows:

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$$

where y_i is the continuous outcome value of subject i, β_0 is intercept, β_j is the coefficient of determinant j and x_{ij} is determinant j of subject i. The unknown β_j can be approximated as follows by minimizing the residual sum of squares.

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\}$$

Correlations between data measures will occur under specific circumstances. This happens when individuals are grouped or when a person is subjected to repeated measurements. In these instances, it is critical to incorporate this connection into the model, as independence is an underlying premise of this model. A mixed model might be employed rather than a linear model when this association exists.

2.6.2 Penalized LR

When there is a multicollinearity problem, penalized regression methods are designed to handle the regression analysis. The penalized regression technique is derived from the least-squares method with a penalty function to identify significant explanatory variables and improve prediction accuracy in linear regression. The ridge regression

(Hoerl and Kennard, 1970), the lasso (Tibshirani, 1996), and the elastic net are all examples of penalized regression (Zou and Trevor, 2005).

2.6.2.1 Lasso regression

The abbreviation lasso refers to the least absolute shrinkage and selection operator. It penalizes the regression model with a term named L1-norm, the total of the absolute coefficients. In the instance of lasso regression, the penalty has the effect of driving the estimation of a few coefficients to be precisely equal to zero in terms of contribution to the model while having a minor impact on others. This means that the lasso can also be used in place of subset selection procedures for lowering the complexity of the model through variable selection. As with ridge regression, choosing an appropriate value for the lasso is crucial.

One obvious advantage of lasso regression over ridge regression is that it provides more straightforward and interpretable results models by utilizing only a subset of the predictors. However, neither ridge regression nor the lasso will consistently outperform the other (Van Wieringen, 2018). The penalty in lasso is equal to the sum of the coefficients' absolute values. When lambda is big enough, lasso decreases the coefficient estimates towards zero, but ridge does not. As a result, lasso conducts variable selection similar to the best subset selection approach. Cross-validation is used to determine the tuning parameter lambda. When lambda is small, the resulting estimates are effectively least squares. As lambda rises, shrinkage happens, allowing zero-valued variables to be discarded. Thus, a significant benefit of the lasso is that it combines shrinkage with variable selection. In this model, the penalty term equals the sum of the absolute weights. This objective function can be used to estimate the lasso unknown β_j:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \left| \beta_j \right| \right\}, \lambda \geq 0$$

Here λ is the size of the shrinking. If λ equals 0, the standard linear regression is recovered.

2.6.2.2 Ridge regression

Ridge regression reduces the size of the regression coefficients, resulting in coefficients close to zero for factors that have a small effect on the outcome. The coefficients are reduced by penalizing the regression model with a term termed L2-norm, equal to the squared coefficients' sum. The penalty amount can be adjusted using a constant named lambda. The penalty term is ignored when lambda equals zero, and ridge regression produces the traditional least-squares coefficients. However, the shrinkage penalty becomes more significant as it climbs to infinity, and the ridge regression coefficients approach zero.

Ridge regression performs better when the outcome is a function of a large number of predictors, each of which has an equal number of equal-sized coefficients. In comparison to conventional least squares regression, ridge regression is strongly influenced by the scale of the predictors. Ridge regression approximates zero by reducing the coefficients to zero. The lasso regression technique provides a workaround for this limitation (Hoerl and Kennard, 1970). Cross-validation approaches can be used to determine which of these two procedures is more appropriate for a given data set.

In ridge regression, we use a lambda tuning parameter determined using cross-validation. The objective is to minimize the fit by reducing the residual sum of squares and applying a shrinkage penalty. The bias remains constant as the tuning parameter increases, but the variance decreases. By minimizing this objective function, the ridge unknown β_j may be calculated.

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}, \lambda \geq 0.$$

2.6.2.3 Elastic net

Zou and Hastie (2005) suggested that the elastic net regression model extend the Lasso by resolving its limitations, particularly variable selection. Furthermore, the elastic net promotes grouping by grouping highly correlated predictors in the model. By contrast, the Lasso algorithm tends to split such collections into subgroups using only the strongest variable. Additionally, the elastic net is favourable when the number of predictors (p) in a data set is greater than the number of observations (n). The Lasso cannot pick more than n predictors in this scenario, whereas the elastic net is.

Elastic nets provide regression models that are penalized according to both the Lasso and ridge methods. This effectively decreases coefficients (as in ridge regression) and zero out some coefficients. Besides defining and selecting a lambda value, elastic nets enable us to tweak the parameter, where 0 corresponds to ridge and 1 to lasso. Expressed, when alpha is set to 0, the penalty function reduces to the L1 (lasso) term; when alpha is set to 1, the L2 (ridge) term is obtained. As such, we can optimize the elastic net by selecting an alpha value between 0 and 1. This effectively shrinks specific coefficients and sets others to zero to facilitate the sparse selection (Friedman *et al.*, 2009). By minimizing this objective function, the elastic net unknowns β_j can be estimated:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum x_{ij}\beta_j \right)^2 + \lambda_1 \sum_{j=1}^{p} \left| \beta_j \right| + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \right\}.$$

2.6.3 Machine learning

2.6.3.1 Random Forest

The disadvantage of multiple linear regression is that it does not fully capture nonlinear interactions between dependent and independent variables. Rather than that, an ensemble method known as RF is employed to predict the outcome. By integrating numerous decision trees, RF is an ensemble approach for predicting the value of an effect. Each tree (model) in the ensemble predicts a new random sample, and the predicted values are summed to give the forest's prediction. The tuning parameter for an RF is the number of predictors randomly picked at each split; this value is denoted by the variables mtry and ntree, which are collectively referred to as hyperparameters (Breiman, 2001). At each split, the algorithm chooses the number of predictors at random.

The hyperparameter named mtry is the number of variables randomly selected as testing conditions at each split of decision trees. Increasing mtry generally improves the model's performance as each node has more options to consider. However, it also decreases the diversity of individual trees.

The hyperparameter called ntree is the number of trees to be planted. The RF model contained 1,000 trees. Cross-validation was utilized to optimize hyperparameters for each model.

2.6.3.2 Neural Network (NN)

NN is a powerful nonlinear regression technique inspired by the brain's operation theories. An intermediary set of unobserved variables is used to model the outcome (called hidden variables or hidden units). In their nature and application, NNs are similar to linear regression models. They are composed of input (independent or predictor variable) and output (dependent or outcome variable) nodes and learn or train (parameter estimation) a model using connection weights, bias weights and cross-entropy. The neurons in the hidden layer communicate exclusively with other neurons and never directly with the user program. NN acquire the ability to perform tasks through inductive learning algorithms that require massive data sets (McCulloch and Pitts, 1943).

The NN model was created and evaluated in this study to predict hospital visit costs associated with chronic illness. Numerous variables have been identified as DRGs. Age-gender group, principal diagnosis, number of ‘diagnoses, number of procedures, discharge status, and length of hospital stay are all included as input variables for the NN model. Using data, we constructed and trained a model based on the multi-layer perceptron topology. The examination of the test data demonstrates that the NN model can accurately estimate the total cost of visits. Our study employs a 3-layer network. An input layer containing n (n = 6) neurons represents the six influencing factors on total hospital visit expenses, a hidden layer containing l (l = 10) neurons, and an output layer containing just 1.

2.6.3.3 Support Vector Regression (SVR)

Support vector machines (SVM) have been widely utilized in supervised learning to solve classification difficulties. Support Vector Regression is based on the same premise as SVM, but it can solve regression issues (Vapnik, 1995). The model built by support vector classification requires only a portion of training data, as the cost function for developing the model ignores training points outside the margin. SVR models are constructed using only a fraction of training data, as the cost function rejects samples with a prediction close to the target. Additionally, all training points are contained within the decision border. SVR aims to find the best-fit hyperplane line and includes the most outstanding data points. SVR is a supervised learning model frequently used in machine learning to solve regression problems. SVR uses nonlinear transformations to generate a set of hyperplanes in a high-dimensional space based on the following function (Zhao and Qi, 2015).

$$f(x) = w.x + b$$

Here $x \in X$ is a vector of the input predictors, $w \in X$ is the weight vector of x, and b is the error that determines the distance of the hyperplane from the original.

SVR minimizes prediction error by reducing the gap between expected and observed output values. As a result, it employs b as a constraint to constrain the magnitude of the normal weight vector.

$$min\frac{1}{2}\|w\|^2$$

2.6.3.4 Xtreme Gradient Boosting (XGBoost)

XGBoost utilizes a gradient-boosting decision tree (GBDT) technique to solve classification and regression issues (Chen and Guestrin, 2016). The greedy approach optimizes the objective function's maximum gain when creating each tree layer. The algorithm's concept is to grow a tree by continually adding trees and performing feature splitting. Each time a new tree is created, the algorithm learns a new function to fit the residual from the previous forecast. Finally, many learners are combined to create the final prediction, which is more accurate than a single one. To address overfitting, XGBoost limits the model's complexity via regularization terms, and objective function optimization computes pseudo residuals using the second derivative of the Taylor expansion loss function (Wu *et al*., 2020).

2.6.4 Bootstrapping

Bootstrapping is a technique for resampling that combines random sampling and replacement (replicating the sampling process). By bootstrapping, the accuracy of a sample estimate is determined. This approach predicts the sample distribution of virtually any statistic using random sampling (Pathak and Rao, 2013). The bootstrap method estimates an estimator's properties by sampling an approximate distribution. When it is assumed that a set of observations originated from a randomly distributed population, it is possible to generate a collection of resamples using replacement and of equal size to the observed data set.

2.6.5 Comparison procedure

The r^2 was used to indicate how well the values fit together compared to the initial values. R-squared has a value between 0 and 1, and the optimal score is 1.0. The greater the value, the more accurate the model. This statistic can be considered the proportion of variance explained by the response (dependent) variable model. It is derived by squaring the correlation coefficient between the observed and anticipated values.

The following metrics were used to evaluate the performance of the regression models: The root mean squared error (RMSE) is defined as the difference between expected and observed/actual values: residuals equal the difference between observed and predicted values. The mean squared error (MSE) is then calculated using the residuals, squared, added together, and divided by the sample size.

A sample size of n is used. Calculate RMSE by multiplying the MSE by the square root of the MSE, which is expressed in the same units as the original data. The RMSE number indicates approximately how far the predicts are (on average) from the actual data.

All graphs, data processing and manipulation, and statistical analysis were performed using R statistical (R Core Team, 2020).

CHAPTER 3

RESULT

For this thesis, two manuscripts were produced. The first manuscripts, entitle "Determinants of Hospital Costs for Management of Chronic-Disease Patients in Southern Thailand" was published in the Journal of Health Science and Medical Research. The second manuscript, entitled "Comparison of Linear, Penalized Linear and Machine Learning Models Predicting Hospital Visit Costs from Chronic Diseases in Thailand" was published in the Informatics in Medicine Unlocked. The full manuscripts are shown in the Appendix. This chapter describes the results from the analyses, which are both included and not included in the manuscripts.

3.1 Article I: Determinants of Hospital Costs for Management of Chronic-Disease Patients in Southern Thailand

3.1.1 Preliminary results

This section starts with a description of hospital visit cost characteristics associated with chronic disease of Southern Thailand in 2016. A total of 18,342 hospital visit costs were from the hospital claim database. The preliminary analysis of hospital visit cost deals with descriptive analysis and characteristics of the study variables. Table 3.1 provides the results obtained from the preliminary analysis of hospital visit costs. For each variable, the descriptive results are presented by numbers of records and percentages.

Table 3.1 summarizes the characteristics of the patients. Male patients represented more than half (55.6%) of all patients, and approximately 57% were aged 50 to 79 years. The median LOS was 3 days. The majority of patients reported LOS of 1-3 days. Respiratory diseases accounted for the greatest percentage (12.7 percent), followed by ischemic heart disease (11.2 percent) and cancers other than liver or lung cancer (11.1 percent). While most patients (32.3 percent) had only one surgery, nearly 62% had between 2 and 5 diagnoses. The median number of procedures was 2 (min=0, max=12), while the median number of diagnoses was 4 (min=1, max=13).

Table 3.1 Demographic and clinical characteristics of study patients

Demographic characteristics	Number	Percent
Gender-age groups		
Male		
0 – 9 years	854	4.7
10 – 19 years	183	1.0
20 – 29 years	272	1.5
30 – 39 years	536	2.9
40 – 49 years	1,258	6.9
50 – 59 years	2,051	11.2
60 – 69 years	2,016	11.0
70 – 79 years	1,952	10.6
80 – 89 years	945	5.2
90+ years	134	0.7
Female		
0 – 9 years	567	3.1
10 – 19 years	163	0.9

Table 3.1 (cont.)

Demographic characteristics	Number	Percent
20 – 29 years	204	1.1
30 – 39 years	519	2.8
40 – 49 years	983	5.4
50 – 59 years	1,483	8.1
60 – 69 years	1,478	8.1
70 – 79 years	1,488	8.1
80 – 89 years	1053	5.7
90+ years	203	1.1
Length of hospital stay (LOS)		
0 day	833	4.5
1 day	3,260	17.8
2 days	3,758	20.5
3 days	2,131	11.6
4 days	1,578	8.6
5 days	1,140	6.2
6 days	765	4.2
7-8 days	1,266	6.9
9-11 days	1,158	6.3
12-15 days	791	4.3

Table 3.1 (cont.)

Demographic characteristics	Number	Percent
16-24 days	890	4.9
25+ days	772	4.2
LOS median (min, max)	3 (0, 346)	
ICD-10 group		
Respiratory diseases	2,332	12.7
Ischemic heart disease	2,046	11.2
Other cancers	2,030	11.1
Genitourinary diseases	1,417	7.7
Digestive diseases	1,338	7.3
Other digestive diseases	1,282	7.0
Stroke	1,289	7.0
Other cardiovascular diseases	1,152	6.3
Endocrine diseases	702	3.8
Ill-defined	460	2.5
Liver cancer	374	2.0
Lung cancer	298	1.6
Other infectious diseases	284	1.5
HIV/AIDS	271	1.5
Muscle and nervous system	266	1.5

Table 3.1 (cont.)

Demographic characteristics	Number	Percent
Tuberculosis	202	1.1
Septicemia	179	1.0
Other	2,420	13.2
Discharge status		
Exited	16,168	88.1
Died without autopsy	1,445	7.9
Died with autopsy	114	0.6
Escaped	561	3.1
Denied treatment	160	0.9
Other	8	0.04
Number of diagnoses (nDiag)		
1	1,138	6.2
2	2,836	15.5
3	3,069	16.7
4	3,168	17.3
5	2,328	12.7
6	1,590	8.7
7	1,209	6.6
8	854	4.7

Table 3.1 (cont.)

Demographic characteristics	Number	Percent
9	627	3.4
10	422	2.3
11	312	1.7
12	220	1.2
13	569	3.1
nDiag: median (min, max)	4 (1, 13)	
Number of procedures (nProc)		
0	2,891	15.18
1	5,934	32.4
2	3,756	20.5
3	1,936	10.6
4	1,160	6.3
5	668	3.6
6	764	4.2
7	424	2.3
8	230	1.3
9	165	0.9
10	125	0.7
11	49	0.3

Table 3.1 (cont.)

Demographic characteristics	Number	Percent
12	240	1.3
nProc: median (min, max)	2 (0, 12)	

Note: ICD -10: International Classification of Diseases version 10, HIV: Human Immunodeficiency Virus

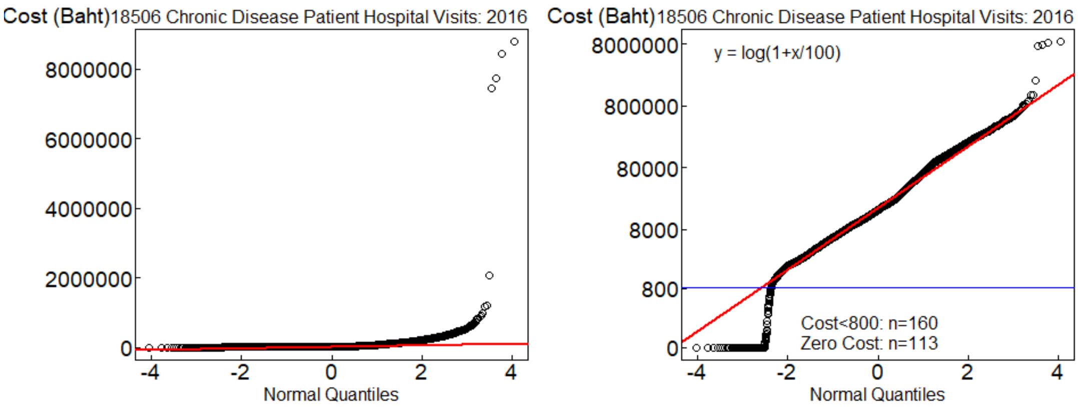


Figure 3.1 Normal quantile-quantile plots of cost (left) and transformed cost (right)

On the left, the quantile-quantile (Q-Q) plot of expenditures for the original sample of 18,506 patients shows a highly skewed distribution with 4 large outliers for visits totalling more than 7 million Baht. However, after applying the log(1+cost/100) transformation, the distribution remained normal, except for small groups at the low and high extremes, as illustrated in Figure 3.1, right-hand plot. As previously stated, expenses less than 800 baht were removed from further analysis, and several linear models were developed. The model included gender-age group, LOS, ICD-10 group, discharge status, nDiag, and nProc.

The results from the model using treatment contrast are shown in Table 3.2. All of the determinants had a significant relationship with hospital costs. Each determinant with the code with the lowest value was automatically chosen as the reference group for the treatment contrast method. However, this study applied the sum contrast to compare each category with the overall mean of hospital cost. Thus, the multiple linear models using sum contrast was created. The coefficients were converted to the hospital cost and illustrated using a 95% CI plot, as shown in Figure 3.2.

Table 3.2 The relationship between hospital cost and predictors from linear regression model

Determinants	Coefficient	Std. Error	P-value
Constant	7.16	0.07	<0.001***
Gender-age groups			
Male			
0- 9 years (ref.)	0.00		
10 – 19 years	0.31	0.05	<0.001***
20 – 29 years	0.49	0.44	<0.001***
30 – 39 years	0.47	0.04	<0.001***
40 – 49 years	0.54	0.03	<0.001***
50 – 59 years	0.53	0.03	<0.001***
60 – 69 years	0.54	0.03	<0.001***
70 – 79 years	0.55	0.03	<0.001***
80 – 89 years	0.43	0.03	<0.001***

Table 3.2 (cont.)

Determinants	Coefficient	Std. Error	P-value
90+ years	0.46	0.06	<0.001***
Female			
0 – 9 years	0.05	0.03	0.131
10 – 19 years	0.32	0.05	<0.001***
20 – 29 years	0.47	0.05	<0.001***
30 – 39 years	0.46	0.04	<0.001***
40 – 49 years	0.48	0.03	<0.001***
50 – 59 years	0.48	0.03	<0.001***
60 – 69 years	0.56	0.03	<0.001***
70 – 79 years	0.46	0.03	<0.001***
80 – 89 years	0.44	0.03	<0.001***
90+ years	0.43	0.05	<0.001***
Length of hospital stay			
0 day (ref.)	0.00		
1 day	0.58	0.03	<0.001***
2 days	0.76	0.02	<0.001***
3 days	0.88	0.03	<0.001***
4 days	1.05	0.03	<0.001***
5 days	1.22	0.03	<0.001***
6 days	1.35	0.03	<0.001***

Table 3.2 (cont.)

Determinants	Coefficient	Std. Error	P-value
7-8 days	1.51	0.03	<0.001***
9-11 days	1.68	0.03	<0.001***
12-15 days	1.87	0.03	<0.001***
16-24 days	2.13	0.03	<0.001***
25+ days	2.56	0.03	<0.001***
ICD 10 group			
Tuberculosis (ref.)	0.00		
Sepsis	0.24	0.03	<0.001***
HIV	0.16	0.03	<0.001***
Other infection	0.23	0.03	<0.001***
Liver cancer	0.54	0.03	<0.001***
Lung cancer	0.01	0.03	0.905
Other digestive	0.46	0.03	<0.001***
Other cancers	0.15	0.03	0.001**
Endocrine diseases	0.16	0.03	<0.001***
Muscle nervous system	0.11	0.03	0.062
Ischemic heart disease	0.83	0.05	<0.001***
Stroke	0.05	0.05	0.251
Other cardiovascular	0.36	0.05	<0.001***
Respiratory diseases	0.22	0.05	<0.001***

Table 3.2 (cont.)

Determinants	Coefficient	Std. Error	P-value
Digestive diseases	0.21	0.05	<0.001***
Genitourinary diseases	0.01	0.05	0.759
Ill-defined diseases	0.13	0.04	0.014*
Other	0.10	0.05	0.019*
Discharge status			
Exited (ref.)	0.00		
Died without autopsy	0.06	0.05	0.229
Died with autopsy	0.17	0.05	0.001**
Escaped	- 0.31	0.22	0.153
Denied treatment	0.18	0.07	0.014*
Other	0.22	0.05	<0.001***
Number of diagnoses			
1 (ref.)	0.00		
2	0.01	0.02	0.518
3	0.06	0.02	0.012*
4	0.08	0.02	<0.001***
5	0.02	0.02	0.356
6	-0.02	0.03	0.467
7	- 0.05	0.03	0.086
8	- 0.09	0.03	0.002**

Table 3.2 (cont.)

Determinants	Coefficient	Std. Error	P-value
9	- 0.06	0.03	0.052
10	- 0.04	0.04	0.277
11	- 0.04	0.04	0.396
12	-0.11	0.05	0.025*
13	- 0.04	0.04	0.322
Number of procedures			
0 (ref.)	0.00		
1	0.50	0.01	<0.001***
2	0.78	0.02	<0.001***
3	1.05	0.02	<0.001***
4	1.28	0.02	<0.001***
5	1.42	0.03	<0.001***
6	1.94	0.03	<0.001***
7	1.89	0.03	<0.001***
8	1.95	0.04	<0.001***
9	1.95	0.05	<0.001***
10	2.08	0.06	<0.001***
11	2.11	0.08	<0.001***

Note: * p<0.05, ** p<0.01, *** p<0.001

The coefficients and confidence intervals for the linear regression were used to predict the hospital costs associated with chronic disease. When the CI is greater than or less than the mean, the mean shows the more or less important group than the overall mean. Gender, age group, LOS, ICD 10 classification, nProc, nDiag, and discharge status significantly affected hospital cost, with LOS and nProc having a higher hospital cost than the overall mean.

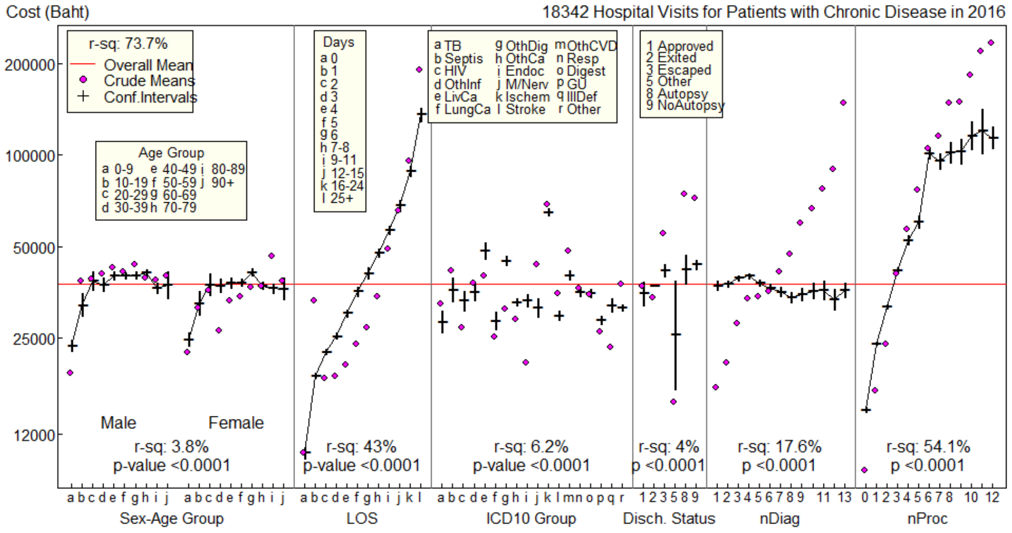


Figure 3.2 95 % CI plot of hospital costs and determinants from multiple linear regression model

Note: Disch. Status = discharge status

Figure 3.2 shows the results from the multiple linear regression model. The r-squares from the simple linear model of each determinant with hospital cost are presented at the bottom of the plot. P-values from multiple linear models are also illustrated at the bottom of the plot. The CI plots demonstrated a prediction of 73.7 percent, with nProc and LOS being the strongest predictors. When nProc and LOS were included, diagnosis (ICD-10 group) and nDiag performed poorly. The crude means for nDiag (circle dots) indicated a good predictor of cost. However, this association disappeared when LOS and nProc were included in the model. As such, it was determined to be a confounding variable. A simple linear model was then fitted for each determinant, and the r^2 values indicated that nProc had the best predictive value (54.1%), followed by LOS (43.0%) and nDiag (17.6%). Thus, nProc and LOS were the only two variables included in the final model, despite all other variables having significant p-values.

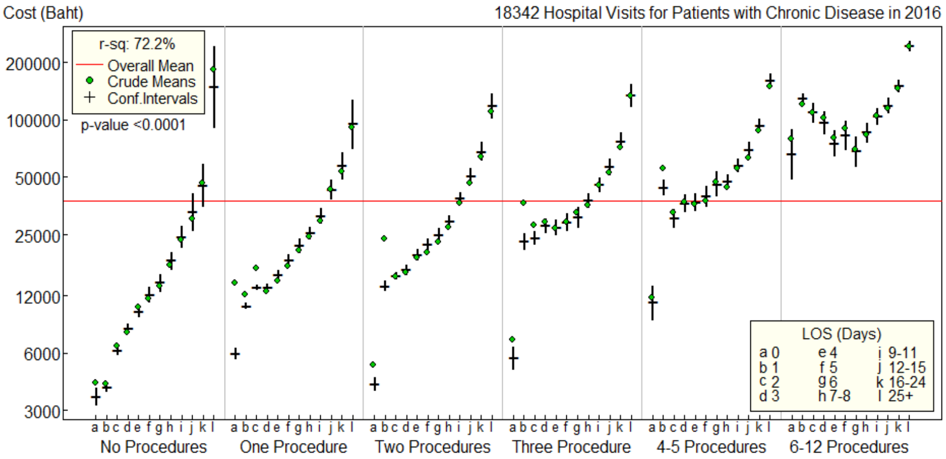


Figure 3.3 The 95% CI plot of the relationship between hospital cost and nProc-LOS

Although just two of the initial six components were preserved in the final model, the r^2 was reduced by only 0.015. Thus, medical costs increased as nProc and LOS increased, except for patients who underwent six to twelve operations during hospitalization.

Boxplots of hospital cost separated by each group of LOS and nProc were created as shown in Figures 3.4-3.8. The green dots in the plot denotes crude means. The red plus sign represent the 95% CI.

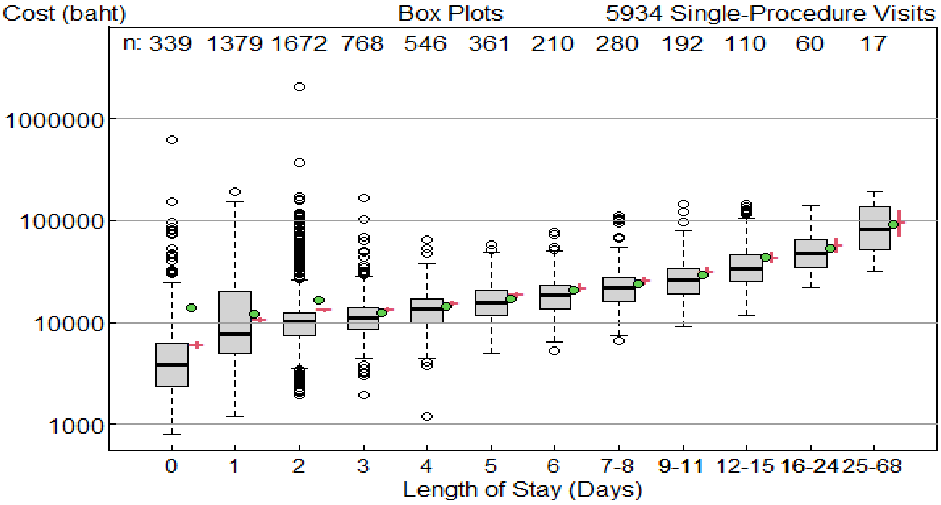


Figure 3.4 Distribution of hospital cost separated by LOS for single procedure visit

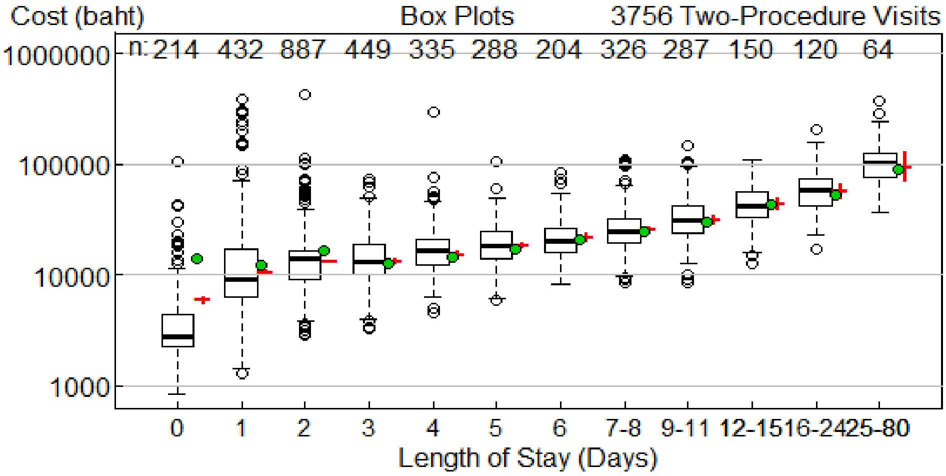


Figure 3.5 Distribution of hospital cost separated by LOS for two procedure visits

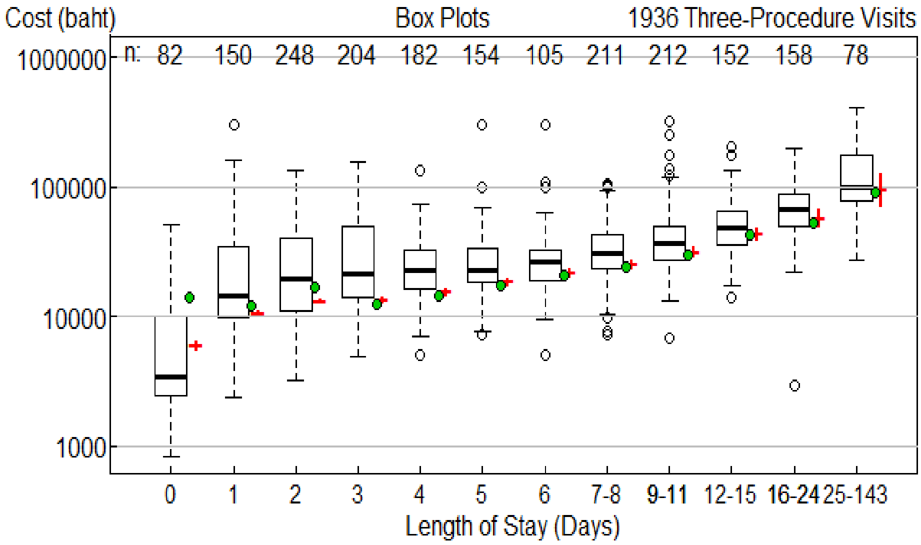


Figure 3.6 Distribution of hospital cost separated by LOS for three procedure visits

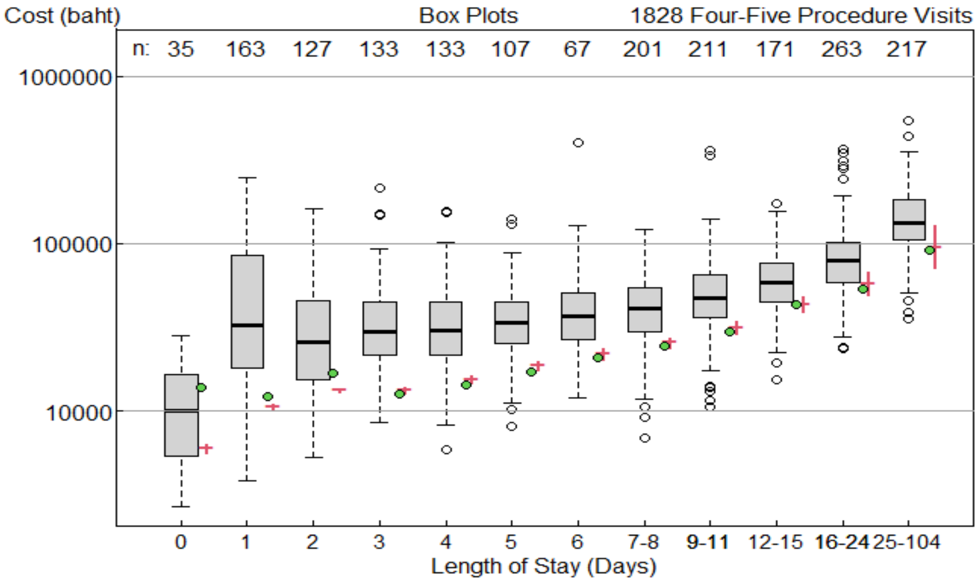


Figure 3.7 Distribution of hospital cost separated by LOS for four-five procedure visits

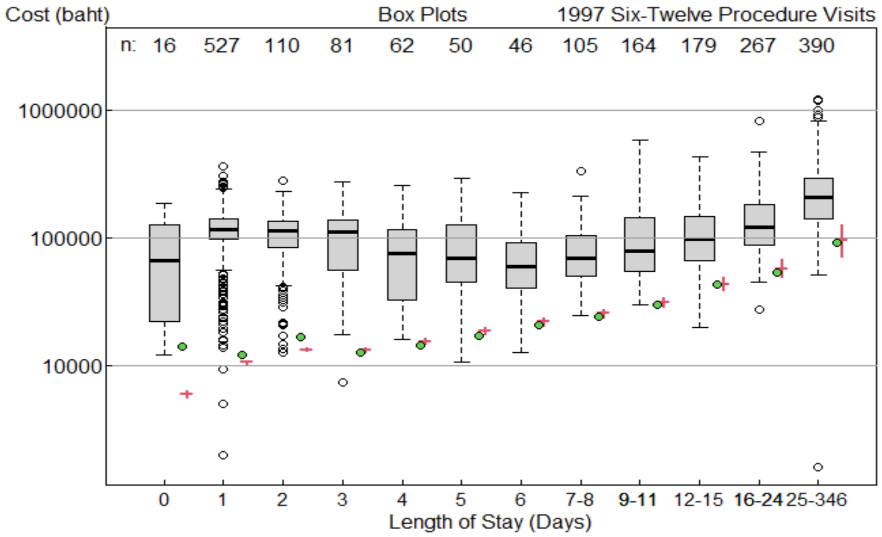


Figure 3.8 Distribution of hospital cost separated by LOS for six-twelve procedure visits

Hospital cost increased by increasing LOS for all nProc except for nProc 6-12, as shown in Figures 3.4-3.8

3.2 Article II: Comparison of Linear, Penalized Linear and Machine Learning Models Predicting Hospital Visit Costs from Chronic Diseases in Thailand

3.2.1 Preliminary results

This study analyzed a total of 18,342 admission records for chronic illnesses. The age groupings were as follows: 0-9, 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, and 90 and above. The natural logarithm of hospital cost in Thai Baht divided by 100 and adding one (1 USD is about 31 THB) per patient visit is the outcome. As previously stated, ICD-10 was categorized into 18 groups.

About 55.6% of patients were males. The majority of them, 57.0 %, were between the ages of 50 and 79. The median number of procedures was 2 (min=0, max=12) while the median number of diagnoses was 4 (min=1, max=13). The median LOS was 3 days. Respiratory diseases, ischemic heart disease and cancers other than liver or lung cancer were found for 12.7%, 11.2% and 11.1%, respectively.

Table 3.3 Prediction performance of each model

Model	Training		Testing	
	RMSE	R^2	RMSE	R^2
LR				
Original	0.5891	74.1	0.6040	73.0
2 times	0.5944	73.5	0.6029	73.1
4 times	0.5940	73.8	0.5953	73.4
Lasso				
Original	0.5892	74.1	0.6041	73.0
2 times	0.5945	73.5	0.6030	73.0
4 times	0.5941	73.8	0.5955	73.4
Ridge				
Original	0.6035	73.1	0.6185	71.8
2 times	0.6088	72.4	0.6150	72.1
4 times	0.6082	72.8	0.6092	72.3
Elastic Net				
Original	0.5892	74.1	0.6041	73.0
2 times	0.5945	73.5	0.6030	73.1
4 times	0.5941	73.8	0.5955	73.4

Table 3.3 (cont.)

Model	Training		Testing	
	RMSE	R^2	RMSE	R^2
SVR				
Original	0.5969	74.2	0.6083	72.8
2 times	0.5711	75.7	0.5811	75.1
4 times	0.5543	77.3	0.5569	76.8
NN				
Original	0.4760	83.1	0.5413	78.4
2 times	0.4774	82.9	0.5181	80.2
4 times	0.4831	82.7	0.4965	81.5
RF				
Original	0.4120	87.8	0.5286	79.5
2 times	0.3449	91.4	0.4194	87.2
4 times	0.3146	92.8	0.3560	90.6
XGBoost				
Original	0.4056	87.8	0.5291	79.3
2 times	0.4006	88.0	0.4526	84.8
4 times	0.4115	87.5	0.4350	85.8

Note: RMSE= Root Mean Square Error, R^2= Coefficient of determination,

LR= Linear Regression, SVR= Support Vector Regression, NN= Neural Network,

RF= Random Forest

Table 3.3 summarizes the predictive performance of each model across all sample sizes. The results for the original dataset indicate that RF and XGBoost outperformed NN in both training and testing datasets, with nearly identical r^2 values, as shown in Table 3.3.

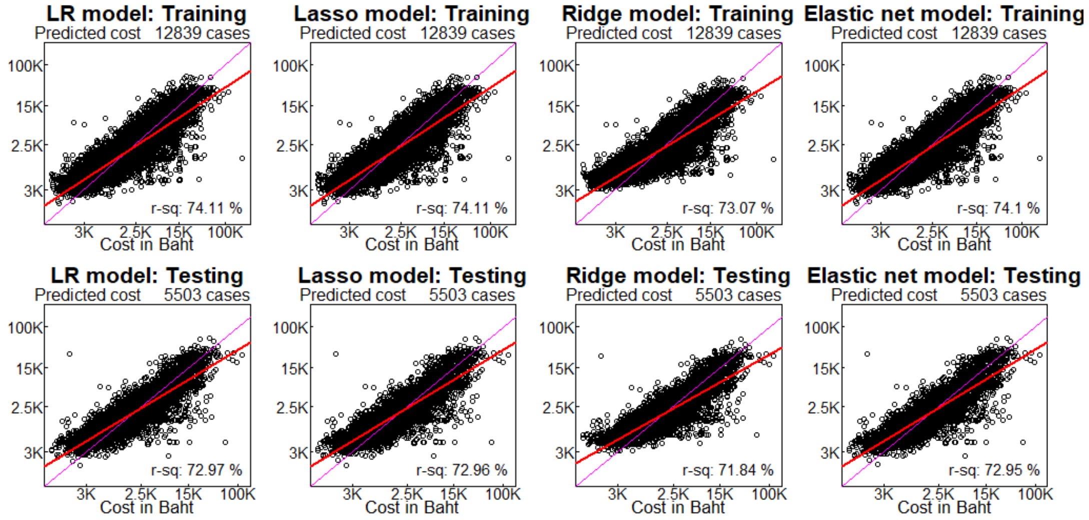


Figure 3.9 (a) shows the results from LR and Penalized LR models using the original hospital visit cost data

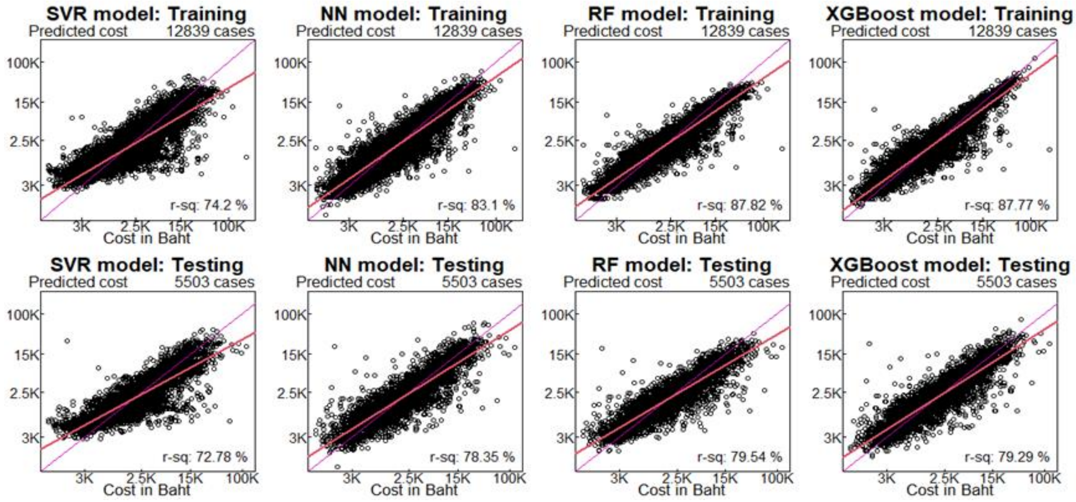


Figure 3.9 (b) shows the results from ML models using the original hospital visit cost data

Figure 3.9 (a) and Figure 3.9 (b) shows the results from LR, lasso, elastic net, and SVR performed similarly in the training and testing datasets, although ridge regression had the lowest r^2 in the testing dataset.

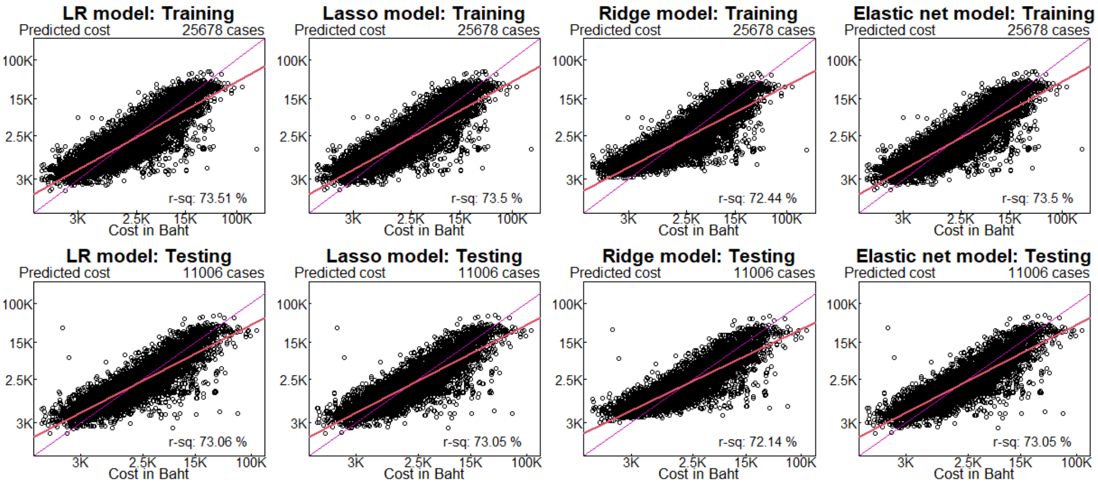


Figure 3.10 (a) shows the results from LR and Penalized LR models using the bootstrap 2 times original hospital visit cost data.

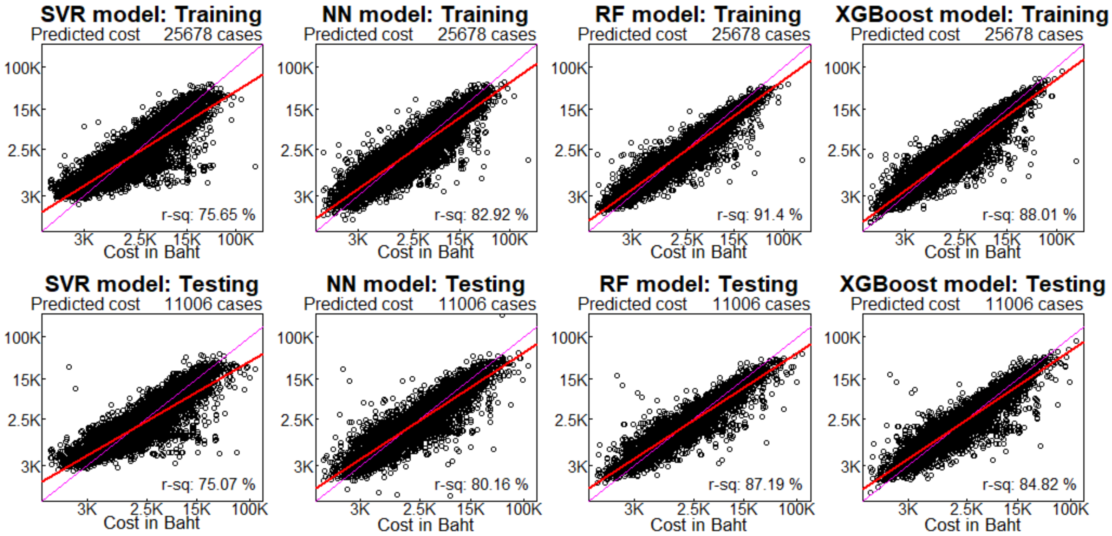


Figure 3.10 (b) shows the results from ML models using the bootstrap 2 times original hospital visit cost data.

As illustrated in Figure 3.10 (a) and Figure 3.10 (b), RF exhibited the best prediction performance after doubling the sample size. The results indicate that RF outperformed all other models in both the training and testing datasets, with r^2 values of 0.914 and 0.872 and RMSE values of 0.3418 and 0.4093, respectively, followed by XGBoost with r^2 values of 0.880 and 0.848 and RMSE values of 0.4006 and 0.4526, NN with r^2 values of 0.829 and 0.802 and RMSE values of 0.4688 and 0.5033, and Ridge regression performed the worst in both training and testing for the doubled dataset, with r^2 values of 0.724 and 0.721 and RMSE values of 0.6088 and 0.6150.

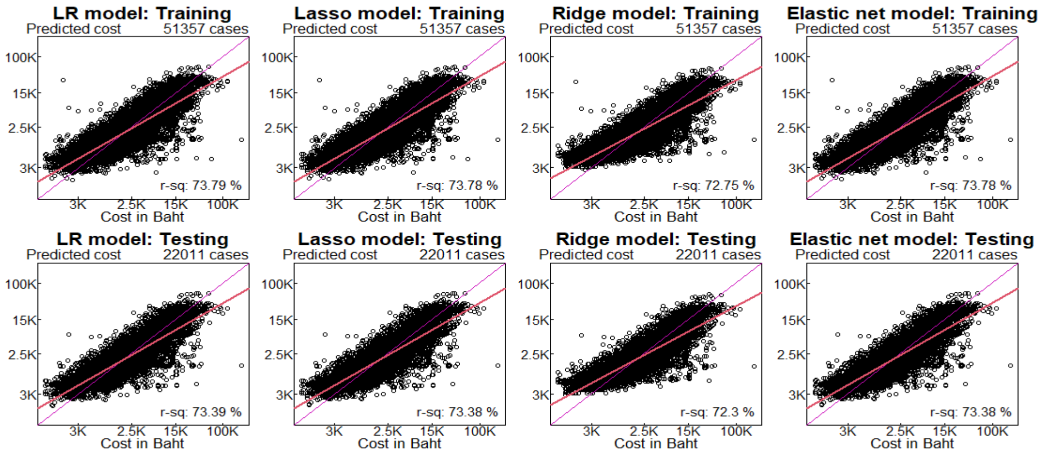


Figure 3.11 (a) shows the results from LR and Penalized LR models using the bootstrap 4 times in hospital visit cost data.

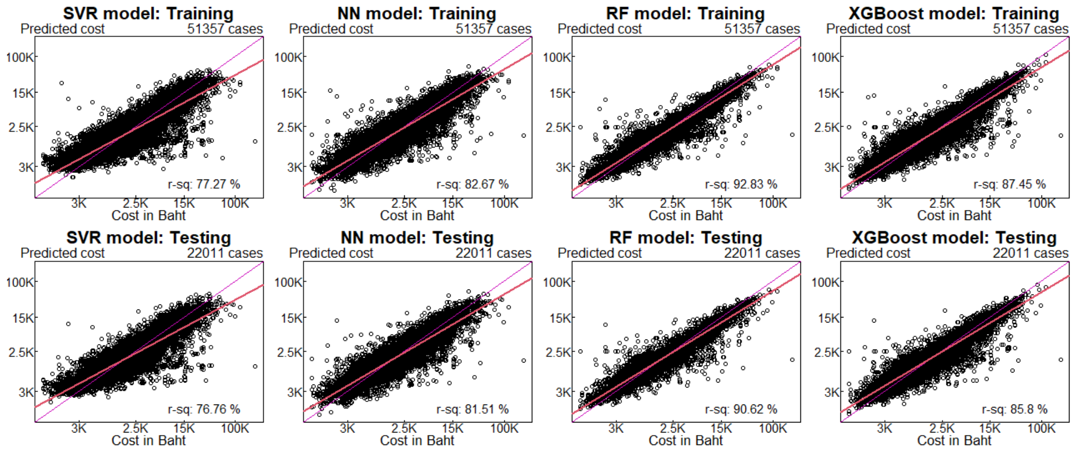


Figure 3.11 (b) shows the results from ML using the bootstrap 4 times in hospital visit cost data.

Even after quadrupling the sample size, the RF retained the greatest prediction ability, as illustrated in Figure 3.11 (a) and Figure 3.11 (b). The results indicate that RF outperformed XGBoost in both the training and testing datasets, with r^2 values of 0.928 and 0.906 and RMSE values of 0.3125 and 0.3542, respectively, followed by NN with r^2 values of 0.875 and 0.858 and RMSE values of 0.4115 and 0.4350 and r^2 values of 0.827 and 0.815 and RMSE values of 0.4784 and 0.4922. SVR's prediction performance was ranked fourth in the four-fold extended dataset, with r^2 values of 0.773 and 0.768 and RMSE values of 0.5507 and 0.55 for training and testing, respectively. Ridge performed poorly in both training and testing, with r^2 values of 0.726 and 0.723 and RMSE values of 0.6081 and 0.6092.

CHAPTER 4

DISCUSSION AND CONCLUSIONS

This chapter summarizes and discusses the results of the thesis and presents the conclusions of the study. The discussions are presented in two parts. The first part explains the determinants of hospital cost of DRGs in Thailand, whiles the second part focuses on the predictive accuracy of various statistical models and machine learning algorithms for predicting hospital cost. Conclusions from the results and recommendations to health care policymakers are also presented in this chapter.

4.1 Determinants of costs for hospitalized chronic disease patients

This thesis has assessed the determinants of costs for hospitalized chronic disease patients. The average hospital cost was 37,644 Baht per visit. All factors that were examined in the DRGs systems were significantly associated with hospital cost, with varying r^2. The number of procedures had the highest r^2 of 53.7 %, followed by the LOS with 42.6%. However, predictors such as gender-age group, principal diagnosis, discharge status and the number of diagnoses had relatively low r^2, less than 20 %, although significantly associated with the outcome. The r^2 is a determinant of the strength of association between predictors and outcome. However, the large sample size of data tends to provide significant results even though the relationship between each determinant and outcome provided low $r^{2,}$ as shown in this study. Thus, we considered LOS and the number of procedures except for more than 6-12 procedures as the main factors influencing hospital cost. The determinants of hospital cost identified in this study have also been documented by other studies (Evans *et al*., 1995; Penberthy, 1999; Silber *et al*., 1999; Philbin *et al*., 2001; Chaikledkaew *et al.*, 2008; Liu *et al*., 2018;

Yuan *et al.,* 2019; Xu *et al.,* 2020). However, this literature reported only p-value without the r^2 values. Our study mainly considers r^2 value rather than the p-value because we analyzed a large dataset. Even a small predicting effect resulted in significant results not relevant to clinical significance. The explanation of longer LOS and higher number of procedures influenced the higher hospital cost is that patients who remain in the hospital for an extended period and have additional procedures spend more hospital resources. An inherent weakness in linear regression is that it cannot adequately capture nonlinear interactions between dependent and independent variables.

In conclusion, LOS and the number of procedures are the significant factors determining hospital costs for patients with chronic diseases. The measures and policies for reducing hospital care costs should focus on these two factors.

4.2 Statistical methods and machine learning algorithms to predict hospital cost

The various statistical methods and ML algorithms applied in this study showed varied predictive accuracy. The performance of LR, penalized LR and ML algorithms are compared. Overall, this study's findings suggest that the RF algorithm outperformed all other algorithms in terms of hospital cost prediction. Also, XGBoost, NN and SVR had higher predictive accuracy than regression models. When the LR, lasso, and elastic net models were enlarged two or four-fold by bootstrapping, they performed nearly identically to the original data. Ridge regression had the lowest performance across all sample sizes. The prediction performance of the standard LR and penalized LR did not change when the sample size increased. These results supported the findings by Mazumdar *et al.* (2020), indiacated that after applying various scenarios of machine learning models the prediction performance are improved when the sample size is doubled. However, similar to other literature (Duncan *et al., 2016*; Rajula *et al., 2020*) on medical cost prediction. Increasing the number of predictors had a more significant effect on the other ML models. This could be explained by the fact that the basic prediction of the LR model is based on the ordinary least square method, which minimizes the prediction error. Therefore, the relationship pattern does not change even with larger sample size, resulting in stable r^2.

In contrast, ML models were developed based on the learning process. Therefore, the ML models can learn if the sample size is large enough. As evident from our analysis, the predictive accuracy increases with SVR, NN, RF and XGBoost models having higher accuracy. The result from our study agrees with the findings by Kulkarni *et al.* (2020), which indicated that ML techniques such as RF and XGBoost have high predictive accuracy with a larger sample size. Our finding is consistent with a study conducted by Lakshmanarao *et al*. (2020) that reported that RF outperformed LR, SVR, DT, and RF in predicting medical expenditures.

Additionally, Kuo *et al.* (2018) discovered that the RF model accurately predicted the profit or loss associated with spinal fusion in Taiwan DRGs. The improved predictive performance of RF over other machine learning and natural language processing models may be explained by the fact that RF produces a forest from numerous decision trees. This is one of the most effective ensemble machine learning techniques for circumventing issues about overfitting data. Our findings, however, contradict those of several previous studies. Seligman *et al.* (2018) investigated the performance of LR, penalized regressions, RF, and NN in predicting health outcomes using socioeconomic determinants of health. They discovered that NN outperformed the other three model types significantly. This could be because the NN model type is the most flexible, allowing for more interactions and non-linear correlations between the determinants and the outcome than other machine learning models typically used with social data. This is one of the most effective ensemble machine learning techniques for addressing the issue of overfitting data. Our findings, however, contradict those of some previous studies.

In conclusion, increased sample sizes had apparent effects on ML models, giving increased r^2 and decreased RMSE, and increasing the sample size did not affect LR and penalized LR. The application of prediction models can depend on sample sizes.

REFERENCES

Ai, C., & Norton, E. C. (2000). Standard errors for the retransformation problem with heteroscedasticity. *Journal Health Economic*, 19(5), 697-718.

Aljunid, S., & Jadoo, S. A. (2018). Factors Influencing the Total Inpatient Pharmacy Cost at a Tertiary Hospital in Malaysia: A Retrospective Study. *INQUIRY: The Journal of Health Care Organization, Provision, and Financing*, 55.

Angstman, K. B., Doganer,Y. C., DeJesus, R. S., & Rohrer, J. E. (2016). Increased medical cost metrics for patients 50 years of age and older in the collaborate care model of treatment for depression. *Psychogeriatrics*, 16(2), 102-106.

Austin, P. C., Ghali, W. A., & Tu, J. V. (2003). A comparison of several regression models for analyzing cost of CABG surgery. *Statistical Methods in Medical Research,* 22, 2799-2815.

Basu, A., Arondekar, B. V., & Rathouz, P. J. (2006). Scale of interest versus scale of estimation: Comparing alternative estimators for the incremental costs of a comorbidity, *Health Economic*, 15, 1091-1107.

Benoit, R. M., & Cohen, J. K. (2001). The relationship between quality and costs: factors that affect the hospital costs of radical prostatectomy, *Prostate Cancer Prostatic Disease*, 4(4), 213-216.

Bernell, S., & Howard, S. W. (2016). Use Your Words Carefully: What Is a Chronic Disease? *Frontline Public Health*, 4, 1592.

Bertsimas, D., Bjarnadóttir, M. V., Kane, M. A., Kryder, J. C., Pandey, R., & Vempala, S. (2008). Algorithmic prediction of health-care costs. *Operations Research*, 56, 1382-1392.

Bhattacharyya, S. K., & Else B. A. (1999). Medical costs of managed care in patients with type 2 diabetes mellitus. *Clinical therapeutics*, 21, 2131–2142.

Bloom, D. E., Chen, S., Kuhu, M., McGovern, M. E., Oxley, L., & Prettner, K. (2017). The economic burden of chronic diseases: Estimates and projections for China, Japan, and South Korea. *The Journal of the Economics of Ageing*, 100163. doi: 10.1016/j.jeoa.2018.09.002.

Boulesteix, A. L., & Schmid, M. (2014). Machine learning versus statistical modeling. *Biometrical Journal*, 56, 588-593.

Bramkamp, M., Radovanovic, D., Erne, P., & Szucs, T. D. (2007). Determinants of costs and the length of stay in acute coronary syndromes: a real-life analysis of more than 10,000 patients. *Cardiovasc Drugs Therapy,* (5),389-398.

Bredenkamp, C., Bales, S., & Kahur, K. (2020). Transition to Diagnosis Related Group (DRG) Payments for Health Lessons from Case Studies. International Development in Focus. *Washington, DC: World Bank*, doi:10.1596/978-1-4648-1521-1528.

Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5-32.

Briestensky, R., & Kljucnikov, A. (2021). The impact of DRG-based management of healthcare facilities on amenable mortality in the European Union. *Problems and Perspectives in Management*, 19(2), 264-275.

Brown, J. B., Pedula, K. L., & Baskt, A. W. (1999). The progressive cost of complications in type 2 diabetes mellitus. *Archives of internal medicine*, 159, 1873–1880.

Busse, R., Geissler, A., Quentin, W., & Wiley, M. (2011). *"Diagnosis Related Groups in Europe: Moving towards Transparency, Efficiency, and Quality in Hospitals?"* European Observatory on Health Systems and Policies.

Bzdok, D., Altman, N., & Krzywinski, M. (2018). Statistics versus machine learning. *Nature Methods*, 15, 233–234.

Cashin, C., O'Dougherty, S., Samyshkin, Y., Katsaga, A., Ibraimova, A., & Kutanov, Y. (2005). *Case-based hospital systems: a step-by-step guide for design and implementation in low- and middle-income countries*. Geneva: Joint United Nations Programme for HIV/AIDS.

Chaikledkaew, U., Pongchareonsuk, P., Chaiyakunapruk, N., & Ongphiphadhanakul, B. (2008). Factors affecting health-care costs and hospitalizations among diabetic patients in Thai public hospitals. *International Society for Pharmacoeconomics and Outcomes Research (ISPOR)*, 11(1), 69-74.

Chapel, J. M., Ritchey, M. D., Zhang, D., & Wang, G. (2017). Prevalence and medical costs of chronic diseases among adult medicaid beneficiaries. *American Journal of Preventive Medicine*, 53, 143-154.

Chen, T., and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 785-794.

Chilingerian, J. (2008). Origins of DRGs in the United States: a technical, political and cultural story. In: Kimberly J, de Pouvourville G, D'Aunno T, editors. *The globalization of managerial innovation in health care,* Cambridge: Cambridge University Press.

Choi, J., Kim, S., Park, H., Jang, S., Kim, T., & Park, E. (2019). Effects of a mandatory DRG payment system in South Korea: Analysis of multi-year nationwide hospital claims data. *BMC Health Services Research*, 19(1), 776.

Collins, J. L., Marks, J. S., & Koplan, J. P. (2009). Chronic disease prevention and control: coming of age at the Centers for Disease Control and Prevention. *Preventing chronic disease*, 6(3), A81.

Conigliani, C., & Tancredi, A. (2009). A Bayesian model averaging approach for cost effectiveness analyses. *Health Economics*, 18 (7), 807-821.

Dans, A., Ng, N., Varghese, C., Tai, E. S., Firestone, R., & Bonita, R. (2011). The rise of chronic non-communicable diseases in Southeast Asia: time for action. *The Lancet*, 377(9766), 680-689.

Ding, R., Jiang, F., Xie, J., & Yu, Y. (2017). Algorithmic prediction of individual diseases. *International Journal of Production Research*, 55(3), 750-768.

Dodd, S., Bassi, A., Bodger, K., & Williamson, P. (2006). A comparison of multivariable regression models to analyse cost data. *Journal of Evaluation in Clinical Practice,* 12, 76-86.

Duan, N., Manning, W. G., Morris, C. N., & Newhouse, J. P. (1983). A comparison of alternative models for the demand for health care. *Journal of Business and Economic Statistics,* 1, 115-126.

Duncan, I., Loginov, M., & Ludkovski, M. (2016). Testing Alternative Regression Frameworks for Predictive Modeling of Health Care Costs. *North American Actuarial Journal*, 20(1), 65-87.

Dureh, N., & Tongkumchum, P. (2019). A comparison of logistic regression and machine learning algorithms applied to zero counts data in contingency tables. *Advances and Applications in Statistics*, 55, 67-76.

Evans, J. H., Hwang, Y., & Nagarajan, N. (1995). Physicians' response to length-of-stay profiling. *Medicare*, 33(11), 1106-1119.

Franzco, R. C. D., & Farmer, L. C. M. (2014). Understanding and checking the assumptions of linear regression: a primer for medical researchers. *Journal of Clinical and Experimental Ophthalmology*, 42, 590-596.

Friedman, J., Latash, M. L., & Zatsiorsky, V. M. (2009). Prehension synergies: A study of digit force adjustments to the continuously varied load force exerted on a partially constrained hand-held object. *Experimental Brain Research,* 197, 1–13.

Gartner, D., Kolisch, R., Neill, D. B., & Padman, R. (2015). Machine Learning Approaches for Early DRG Classification and Resource Allocation. *INFORMS Journal on Computing*, 27(4), 718–734.

Gertman, P. M., & Lowenstein, S. (1984). A research paradigm for severity of illness: issues for the diagnosis- related group system. *Health Care Financing Review*, 12, 79-90.

Gilleskie, D. B., & Mroz, T. A. (2004). A flexible approach for estimating the effects of covariates on health expenditures. *Journal of Health Economics*, 23, 391-418.

Glasgow, R. E., Orleans, C. T., & Wagner, E. H. (2001). Does the chronic care model serve also as a template for improving prevention? *The Milbank Quarterly*, 79, 579-612.

Glynn, L., Valderas, J., Healy, P., Burke, E., Newell, J., Gillespie, P., & Murphy, A. (2011). The prevalence of multimorbidity in primary care and its effect on health care utilization and cost. *Family Practice*, 28(5), 516-523.

Gordon, S. C., Pockros, P. J., Terrault, N. A., Hoop, R. S., Buikema, A., Nerenz, D., & Hamzeh, F. M. (2013). Impact of disease severity on healthcare costs in patients with chronic hepatitis C (CHC) virus infection. *Hepatology*, 56(5):1651-60.

Gregori, D., Petrinco, M., Bo, S., Desideri, A., Merletti, F., & Pagano, E. (2011). Regression models for analyzing costs and their determinants in health care: an introductory review. *International Journal for Quality in Health Care*, 23, (3), 331–341.

Guo, J. J., Gibson, J. T., Gropper, D. M., Oswald, S. L., & Barker, K. N. (1998). Empiric investigation on direct costs-of-illness and healthcare utilization of Medicaid patients with diabetes mellitus. *American Journal of Managed Care*, 4, 1433–1446.

Hamada, H., Sekimoto, M., & Imanaka, Y. (2012). Effects of the per diem prospective payment system with DRG-like grouping system (DPC/PDPS) on resource usage and healthcare quality in Japan. *Health policy*, 107(2-3), 194-201.

Hanafy, M., & Mahmoud, O. (2021). Predict Health Insurance Cost by using Machine Learning and DNN Regression Models. *International Journal of Innovative Technology and Exploring Engineering,* 10(3), 137-143.

Hansen, L. (2016). The importance of epidemiological predictors for healthcare costs for chronic patients: A case study on osteoporotic fracture patients. Aalborg Universitetsforlag. Ph.d.-serien for Det Samfundsvidenskabelige Fakultet, Aalborg Universitet, doi: 10.5278/vbn.phd.socsci.00049.

Hendriks, M. E., Kundu. P., Boers, A. C., Bolarinwa, O. A., Te Pas, M. J., Akande, T. M., …& Swan Tan, S. (2014). Step-by-step guideline for disease-specific costing studies in low- and middle-income countries: a mixed methodology. *Globle Health Action*, 28(7), 23573.

Hill, S. C., & Miller, G. E. (2009). Health expenditure estimation and functional form: applications of the generalized Gamma and extended estimating equations models. *Health Economics*. (In press).

Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics,* 12(1), 55–67.

Holman, H. R. (2020). The relation of the chronic disease epidemic to the health care crisis. *ACR Open Rheumatology*, 2 (3), 167- 173.

Kan, H. J., Kharrazi, H., Chang, H. Y., Bodycombe, D., Lemke, K., & Weiner, J. P. (2019). Exploring the use of machine learning for risk adjustment: A comparison of standard and penalized linear regression models in predicting health care costs in older adults. *PLOS ONE*, 14, 1-13.

Kankeu, H. T., Saksena, P., Xu, K., & Evans, D. B. (2013). The financial burden from non-communicable diseases in low-and middle-income countries: a literature review. *Health Research Policy and Systems*, 11(1), 31.

Kerr, E., Heisler, M., Krein, S., Kabeto, M., Langa, K., Weir, D., & Piette, J. (2007). Beyond Comorbidity Counts: How Do Comorbidity Type and Severity Influence Diabetes Patients' Treatment Priorities and Self-Management? *Journal of General Internal Medicine*, 22(12), 1635-1640.

Kim, Y. K., Stoskopf, C. H., & Saundra, H. G. (2001). Factors Affecting Total Hospital Charges and Utilization for South Carolina Inpatients with HIV/AIDS in 1994–1996. *AIDS Patient Care and STDs*, 15(5), 277-287.

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, 13, 8-17.

Krop, J. S., Powe, N. R., Weller, W. E., Shaffe,r T. J., Saudek, C. D, & Anderson G. F. (1998). Patterns of expenditures and use of services among older adults with diabetes: Implications for the transition to capitated managed care. *Diabetes care*, 21(5), 747-752.

Krop, J. S., Saudek, C. D., Weller, W. E., Powe, N. R., Shaffer, T., & Anderson, G. F. (1999). Predicting expenditures for Medicare beneficiaries with diabetes. A prospective cohort study from 1994 to 1996. *Diabetes Care*, 22(10), 1660-1666.

Kulkarni, S., Ambekar, S. S., & Hudnurkar, M. (2020). Predicting the inpatient hospital cost using a machine learning approach. *International Journal of Innovation Science*, 13(1), 87-104.

Kuo, C. Y., Yu, L. C., Chen, H. C., & Chan, C. L. (2018). Comparison of models for the prediction of medical costs of spinal fusion in Taiwan Diagnosis-Related Groups by machine learning algorithms. *Journal of Healthcare Informatics Research* ,24, 29-37.

Lakshmanarao, A., Koppireddy, C. S., & Kumar, G. V. (2020). Prediction of medical costs using regression algorithms*. Journal of Information and Computational Science*, 10, 751-57.

Lee, S. M., Kang, J. O., & Suh, Y. M. (2004). Comparison of hospital charge prediction models for colorectal cancer patients: neural network vs decision tree models. *Journal of Korean Medical Science*, 19, 677- 681.

Lim, A., Taufik, M. R., Tongkumchum, P., & Dureh, N. (2020). Comparison of different supervised machine learning algorithms for the prediction of tuberculosis mortality. *Advances and Applications in Statistics*, 52, 185-201.

Lin, P. J., Pope, E., & Zhou, F. L. (2018). Comorbidity Type and Health Care Costs in Type 2 Diabetes: A Retrospective Claims Database Analysis. *Diabetes Therapy*. 9(5), 1907-1918.

Liu, X., Kong, D., Lian, H., Zhao, X., Zhao, Y., Xu, Q., … & Fan, Z. (2018). Distribution and predictors of hospital charges for haemorrhagic stroke patients in Beijing, China, March 2012 to February 2015: a retrospective study. *BMJ Open,* 8: e017693. doi:10.1136/ bmjopen-2017-017693.

Malehi, A. S., Pourmotahari, F., & Angali, K. A. (2015). Statistical models for the analysis of skewed healthcare cost data: a simulation study. *Health Economics Review*, 5, 11. doi: 10.1186/s13561-015-0045-7.

Manning, W.G., & Mullahy, J. (2001). Estimating log models: to transform or not to transform? *Journal of Health Economics*. 20, 461-494.

Mathauer, I., & Wittenbecher, F. (2013). Hospital payment systems based on diagnosis-related groups: experiences in low- and middle-income countries. *Bulletin of World Health Organization*, 91, 746-756.

Mazumdar, M., Lin, J. Y. J., Zhang, W., Li, L., Liu, M., Dharmarajan, K., … & Hu, L. (2020). Comparison of statistical and machine learning models for healthcare cost data: a simulation study motivated by Oncology Care Model (OCM) data. *BMC Health Services Research*, 20, 350. doi: 10.1186/s12913-020-05148-y

McCulloch, W. S., & Pitts, W. A. (1943). Logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5, 115-133.

Meetoo, D. (2008). Chronic diseases: the silent global epidemic. *British Journal of Nursing*, 17, 1320-1325.

Mihailovic, N., Kocic, S., & Jakovljevic, M. 2016. Review of diagnosis-related group-based financing of hospital care. *Health Services Research and Managerial Epidemiology*, 3, 1-8.

Mihaylova, B., Briggs, A., O'Hagan, A., & Thompson, S. G. 2011. Review of statistical methods for analysing healthcare resources and costs. *Health economics*, *20*(8), 897–916.

Miranda, J. J., Kinra1, S., Casas, J. P., Smith, G. D., & Ebrahim, S. (2008). Non-communicable diseases in low- and middle-income countries: context, determinants and health policy. *Tropical Medicine and International Health*, 13, 1225-1234.

Motte, S., Melot, C., Pierdomenico, L., Martins, D., Leclercq, P., & Pirson, M. (2015). Predictors of costs from the hospital perspective of primary pulmonary embolism. *European Respiratory Journal*, 47(1), 203-211.

Muremyi, R., Francois, N., Ignace, K., Joseph, N., & Haughton, D. (2019). Comparison of Machine Learning Algorithms for Predicting the Out of Pocket Medical Expenditures in Rwanda. *Journal of Health and Medical Research*, 1(1), 32-41

Nelson-Williams, H., Gani, F., Kilic, A., Spolverato, G., Kim, Y., Wagner, D., … & Pawlik, T.M. (2016). Factors Associated with Interhospital Variability in Inpatient Costs of Liver and Pancreatic Resections. *JAMA Surgery,* 151(2), 155-63.

Or, Z. (2014). Implementation of DRG Payment in France: Issues and recent developments. *Health Policy*, 117(2), 146-150.

Owens, G. M. (2008). Gender differences in health care expenditures, resource utilization, and quality of care. *Journal of Managed Care Pharmacy*, 14(3), 2-6.

Panay, B., Baloian, N., Pino, J. A., Peñafiel, S., Sanson, H., and Bersano, N. (2019). Predicting health care costs using evidence regression. *Proceedings,* 31, 74.

Pathak, P., & Rao, C. R. (2013). The sequential bootstrap. In Handbook of Statistics. *Elsevier*, 31, 2-18.

Patil, P. A., & Salunkhe, A. (2020). Comparative analysis of construction cost estimation using artificial neural networks. *Journal of Xidian University*, 14, 1287-1305.

Peltola, M., & Quentin, W. (2013). Diagnosis-related groups for stroke in Europe: patient classification and hospital reimbursement in 11 countries. *Cerebrovascular Disease*, 35(2), 113-123.

Penberthy, L., Retchin, S. M., McDonald, M. K., McClish, D. K., Desch, C. E., & Riley, G. F. (1999). Predictors of medicare costs in elderly beneficiaries with breast, colorectal, lung, or prostate cancer. *Health Care Management Science*, 2, 149-60.

Perelman, J., & Closon,M. C. (2008). Hospital response to prospective financing of in-patient days: The Belgian case. *Health policy (Amsterdam, Netherlands),* 84, 200-209.

Philbin, E. F., Mccullough, A. P., DEC, W. G., & Disalvo, T. G. (2001). Length of Stay and Procedure Utilization Are the Major Determinants of Hospital Charges for Heart Failure. *Clinical Cardiology*, 24, 56-62.

Pongpirul, K., Walker, D. G., Rahman, H., & Robinson, C. (2011). DRG coding practice: a nationwide hospital survey in Thailand. *BMC Health Services Research*, 11, 290.

Povak, N. A., Hessburg, P. F., McDonnell, T. C., Reynolds, K. M., Sullivan, T. J., & Salter, R. B. (2014). Machine learning and linear regression models to predict catchment-level base cation weathering rates across the southern Appalachian Mountain region, USA. *Water Resource Research*, 50, 2798-814.

Pritchard, D., Petrilla, A., Hallinan, S., Taylor, D. H., Schabert, V., and Dubois, R. (2016). What Contributes Most to High Health Care Costs? Health Care Spending in High Resource Patients. *Journal of Managed Care and Specialty Pharmacy*, 22(2), 102-109.

R Core Team R. (2020). A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing.

Rajula, H. S. R., Verlato, G., Manchia, M., Antonucci, N., & Fanos, V. (2020). Comparison of conventional statistical methods with machine learning in medicine: diagnosis, drug development, and treatment. *Medicina*, 56, 455.

Sakunphanit, T. (2015). Universal health care coverage through pluralistic approaches: experience from Thailand. Bangkok: National Health Security Office.

Sav, A., King, M., Whitty, J., Kendell, E., Mcmillan, S., Kelly, F., Hunter, B., & Wheeler, A. (2015). Burden of treatment for chronic illness: a concept analysis and review of the literature. *Health Expectations*, 18(3), 312-324.

Scheller-Kreinsen, D., Geissler, A., & Busse, R. (2009). The ABC of DRGs. *The Health Policy Euro Observer*, 11(4), 1-5.

Scheller-Kreinsen, D., Geissler, A., & Busse, R. (2013). Breast cancer surgery and diagnosis-related groups (DRGs): Patient classification and hospital reimbursement in 11 European countries. *The Breast*, 22(5), 723-732.

Schreyögg, J., Stargardt, T., Tiemann, O., & Busse, R. (2006). Methods to determine reimbursement rates for diagnosis related groups (DRG): A comparison of nine European countries. *Health Care Management Science*, 9, 215-213.

Seligman, B., Tuljapurkar, S., & Rehkopf, D. (2018). Machine learning approaches to the social determinants of health in the health and retirement study. *SSM Population Health*, 4, 95-99.

Silber, J. H., Gleeson, S. P., & Zhao, H. (1999). The Influence of Chronic Disease on Resource Utilization in Common Acute Pediatric Conditions: Financial Concerns for Children's Hospitals. *Archives of Pediatrics and Adolescent Medicine*, 153 (2), 69–79.

Slabaugh, L. S., Curtis, B. H., Clore, G., Fu, H., & Schuster, D. P. (2015). Factors associated with increased healthcare costs in Medicare Advantage patients with type 2 diabetes enrolled in a large representative health insurance plan in the US. *Journal of Medical Economics,* 18, 106-12.

Sushmita, S., Khulbe, G., Hasan, A., Newman, S., Ravindra, P., Roy, S. B, Cock, M. D, & Teredesai, A. (2016). Predicting 30-day risk and cost of "all-cause" hospital readmissions. The Workshops of the Thirtieth AAAI Conference on Artificial Intelligence Expanding the Boundaries of Health Informatics Using AI: Technical Report WS-16-08.

Sushmita, S., Newman, S., Marquardt, J., Marquardt, J., Prasad, V., & De Cock, M. (2015). Population cost prediction on public healthcare datasets. *Proceedings of the 5th International Conference on Digital Health*, 87-94.

Sweety, B. E., Srimathi, H., & Bagavandas, M. (2019). A Survey of machine learning algorithms in health care. *International journal of scientific and technology research*, 8, 2288-2291.

Swierkowski, P., & Barnett, A. (2018). Identification of hospital cost drivers using sparse group lasso. *PLoS One*, 13(10), e0204300.

Tangcharoensathien, V., Witthayapipopsakul, W., Panichkriangkrai, W., Patcharanarumol, W., & Mills, A. (2018). Health systems development in Thailand: a solid platform for successful implementation of universal health coverage. *Lancet*, 24(391), 1205-1223.

Thorpe, K.E., & Philyaw, M. (2012). The medicalization of chronic disease and costs. *Annual Review of Public Health*, 33, 409-ภ23.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, 58, 267-288.

Toxvaerd, C., Benthien, K., Andreasen, A., Osler, M., & Johansen, N. (2019). Chronic Diseases in High-Cost Users of Hospital, Primary Care, and Prescription Medication in the Capital Region of Denmark. *Journal of General Internal Medicine*, 34(11), 2421-2426.

Van Wieringen, W. N. (2018), Lecture notes on ridge regression, arXiv:1509.09169.

Vapnik, V. (1995). The Nature of Statistical Learning Theory. *Springer*: New York.

Wammes, J. J. G., van der Wees, P. J., & Tanke, M.A.C. (2018). Systematic review of high-cost patients' characteristics and healthcare utilisation. *BMJ Open*, 8, e 023113.

Warren, L. J., Yabroff, K. R., Meekins, A., Topor, M., Lamont, E. B., & Brown, M. L. (2008). Evaluation of Trends in the Cost of Initial Cancer Treatment. *Journal of the National Cancer Institute,* 100, 888 – 897.

World Health Organization (WHO). (2019). World Bank. Global monitoring report on financial protection in health 2019. *World Health Organization and International Bank for Reconstruction and Development / The World Bank; 2019*. Licence: CC BY-NC-SA 3.0 IGO.

Wu, S.W., Pan, Q., & Chen, T. (2020). Research on diagnosis-related group grouping of inpatient medical expenditure in colorectal cancer patients based on a decision tree model. *World Journal of Clinical Cases,* 8(12), 2484-2493.

Xu, Z., Xue, C., Zhao, F., Hu, C., Wu, Y., & Zhang, L. (2020). Hospitalization Costs and Length of Stay in Chinese Naval Hospitals Between 2008 and 2016 Based on Influencing Factors: A Longitudinal Comparison, *Military Medicine,* 185, 15(1). e282-e289.

Yang, C., Delcher, C., Shenkman, E., & Ranka, S. (2018). Machine learning approaches for predicting high-cost high need patient expenditures in health care. *BioMedical Engineering online*, 17, 82-118.

Yuan, S., Liu, W., Wei, F., Zhang, H., Wang, S., Zhu, W., & Ma, J. (2019). Impacts of Hospital Payment Based on Diagnosis Related Groups (DRGs) with Global Budget on Resource Use and Quality of Care: A Case Study in China. *Iranian Journal of Public Health*, 48(2), 238-246

Zhao, D., & Qi, L. (2015). Prediction of maximum power of PV system based on SVR algorithm. *Jilin Institute* of *Chemical Technology*, 32, 89-94.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B*. 67, 301–320.

APPENDIX

Appendix I Article

Determinants of Hospital Costs for Management of Chronic-Disease Patients in Southern Thailand

Original Article

JHSMR Journal of Health Science and Medical Research

Determinants of Hospital Costs for Management of Chronic-Disease Patients in Southern Thailand

Wichayaporn Thongpeth, M.N.S.[1], Apiradee Lim, Ph.D.[1], Sunee Kraonual, M.N.S.[1],
Akemat Wongpairin, M.P.H.[1], Thaworn Thongpeth, M.D.[2]

[1]Department of Mathematics and Computer Science, Faculty of Science and Technology, Prince of Songkla University, Pattani Campus, Pattani 94000, Thailand.

[2]Orthopedic Surgery and Preventive Medicine, Suratthani Hospital, Mueang, Surat Thani 84000, Thailand.

Received 15 June 2020 • Revised 12 September 2020 • Accepted 18 September 2020 • Published online 3 March 2021

Abstract:

Objective: Diagnosis-related groups (DRGs) are the main mechanism for assessing payments for medical treatment. This study aimed to analyze the determinants of costs for chronic-disease patient visits in a major public hospital.

Material and Methods: Hospital cost data available from the hospital database relating to claims made to the Thailand Health Security Office were obtained from a major tertiary hospital for all such patients admitted and discharged in 2016. Linear regression models were created to predict the cost based on several determinants including age and gender, primary diagnosis, number of diagnoses, length of stay, number of procedures, and discharge status.

Results: Only length of stay in hospital and number of procedures were significant predictors of the total hospital costs.

Conclusion: It thus appears that just a combination of these two factors might be a better measure of the true hospital visit costs for patients with chronic disease than DRGs.

Keywords: chronic disease, diagnosis-related groups, hospital costs, length of hospital stay, number of procedures

Contact: Apiradee Lim, Ph.D.
Department of Mathematics and Computer Science, Faculty of Science and Technology,
Prince of Songkla University, Pattani Campus, Pattani 94000, Thailand.
E-mail: apiradee.s@psu.ac.th, api_45@hotmail.com

J Health Sci Med Res 2021;39(4):313-320
doi: 10.31584/jhsmr.2021787
www.jhsmr.org

© 2021 JHSMR. Hosting by Prince of Songkla University. All rights reserved.
This is an open access article under the CC BY-NC-ND license
(http://www.jhsmr.org/index.php/jhsmr/about/editorialPolicies#openAccessPolicy).

Introduction

Health care costs in most countries are mainly determined by the diagnosis-related group (DRG) system, which was initiated in 1983.[1] DRG is a system of classifying patients into groups by standardized prospective payments to hospitals which generally cover all charges associated with an inpatient stay from the time of admission to discharge.[2] The assignment of a patient to a DRG depends on principal diagnosis, secondary diagnosis, surgical procedures performed, comorbidities and complications, patient age and sex and discharge status.[3] The purposes of using the DRG system are cost containment, improving the efficiency, transparency and fairness of funding and quality, and supporting the management of hospitals.[4] Hospitals in most developed countries have introduced DRG as a tool for assessing reimbursement over the past 30 years.[5]

Mathauer and Wittenbecher[6] recommended that the system should be employed to assess hospital costs in low- and middle-income countries with limited resources, even though the main factors for determining the hospital costs might not be the same as those in developed countries. Thus, DRG-based payments in such countries need to be assessed, which might help to improve the efficiency, equity and quality of health services. Globally, DRGs were originally used to calculate reimbursement for hospitals for acute inpatient care but are now also used to assess charges for chronic inpatient care.[6]

In Thailand, health care costs are increasing and one of the main factors influencing this increase is chronic disease, defined as a disease lasting three months or more and generally incapable of being prevented by vaccines or cured by medication.[7] Therefore, it is useful to investigate health care costs among patients suffering from chronic diseases in Thailand

The National Health Security Office (NHSO) of Thailand is the main health-care purchaser in the country and covers 76.0% of the population using the Universal Coverage Scheme which is tax-financed and transfers pooled funds to health providers.[8] The DRG system has been used by the NHSO for over a decade but several issues still exist and need to be solved. Also currently, many hospitals are facing the problem of incurring higher medical care costs than the reimbursement they receive from the NHSO, resulting in hospital financial crises. One possible reason for the widespread financial crises in Thai hospitals could be that the current DRGs might not reflect the real cost of medical care. Identifying the significant factors which influence hospital costs is useful for policy makers in allocating equitable and efficient reimbursement to health providers but no such study of costs has been done recently, thus the system does not have adequate information for informed analyses of the current situation. Therefore, this study aimed to analyze the determinants of costs for visits by patients with chronic diseases to a major public hospital.

Material and Methods

The costs of all hospital visits by patients with chronic diseases including admission and departure dates, age, gender, discharge status, principal and up to 12 secondary diagnoses, up to 12 treatment procedures, and the total visit costs were obtained from the Thailand NHSO. A total of 18,506 records of hospital visits in 2016 by patients with chronic diseases to Surat Thani regional hospital were included in this study. The minimum possible health care cost per visit for a patient related to a chronic disease is 800 Baht. Therefore, following the initial descriptive analysis of the entire sample, all patient visits incurring medical costs of less than 800 Baht or approximately 23 United States dollars (USD) (160 records) and cost more than 7 million (201,250 USD) with one day or less of hospital stay were excluded from further analysis, resulting in 18,342 qualifying records, which were analyzed in this study.

In this study, the main outcome variable was the total cost in Baht per patient visit which included outpatient visit and admission. The determinants were gender, age group, length of hospital stay, International Classification of Diseases version 10 (ICD-10) diagnosis, discharge status, number of diagnoses or comorbidities and complications (nDiag) and number of procedures (nProc). Gender was classified as male or female. Age group was divided into ten groups with 10-year intervals: 0-9, 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89 and 90 and older. Length of hospital stay (LOS) was grouped into 12 groups: 0, 1, 2, 3, 4, 5, 6, 7-8, 9-11, 12-15, 16-24 and 25 or more days. The ICD-10 diagnoses were classified into 18 groups: tuberculosis, sepsis, human immunodeficiency virus (HIV), other infectious diseases, liver cancer, lung cancer, other digestive diseases, other cancers, endocrine diseases, muscle and nervous system diseases, ischemic heart disease, stroke, other cardiovascular diseases, respiratory diseases, digestive diseases, genitourinary diseases, ill-defined diseases, and other diseases. Discharge status was divided into six groups: approved, exited, escaped, other, death with autopsy and death with no autopsy. The number of diagnoses ranged from 1 to 13 and the number of procedures ranged from 0 to 12.

Data cleaning was performed in order to detect and correct coding errors, missing values, outliers and the duplication of records before the statistical analysis was performed. No coding errors and duplicated records were found. There were 4 outliers with medical cost more than 7 million Baht and these records were excluded from the study.

Descriptive statistical analysis was performed to summarize the characteristics of each variable. Medical costs were transformed using natural logarithms. Normal quantile plots were used to depict the distribution of both non-transformed and transformed costs. In order to eliminate the interaction effect of between gender and age group, these two variables were combined to form a new variable called gender-age group with 10 categories. Interactions between other variables were not found and some variables could not be tested due to their small sample size when the variables were combined. Multiple linear regression was used to investigate the relationships between cost and the various determinants. The coefficients and standard errors from the model were converted into cost in Baht and 95% confidence interval (CI) plots were created to illustrate the results from the multivariate analysis. Only significant factors were included in the final model and the results from this model were also illustrated using 95% CI plots. All the statistical analyses were conducted and graphical displays created using the R program, version 3.1.3.[9]

The authors determined that this clinical investigation required Institutional Review Board/Ethics Committee review and approval, and the resulting protocol/approval number was 61/2019.

Results

The summarized patient characteristics are shown in Table 1. More than half of the patients were males (55.6%). About 57.0 % of the patients were aged between 50 and 79 years. Half of the patients had LOS ranging from 1-3 days. Respiratory diseases were found to account for the highest percentage (12.7%) followed by ischemic heart disease (11.2%) and cancers other than liver or lung cancer (11.1%). While the largest group of patients (32.4%), had only one procedure, approximately 62.2% of the patients had from 2-5 multiple diagnoses.

The quantile-quantile (Q-Q) plot of costs for the entire initial sample of 18,506 patients in the left-hand plot of Figure 1 shows a very skewed distribution with four large outliers for visits costing more than 7 million Baht. However, after transformation based on log(1+cost/100), the distribution was found to be normal, apart from small groups at low and high extremes as shown in Figure 1, right-hand plot.

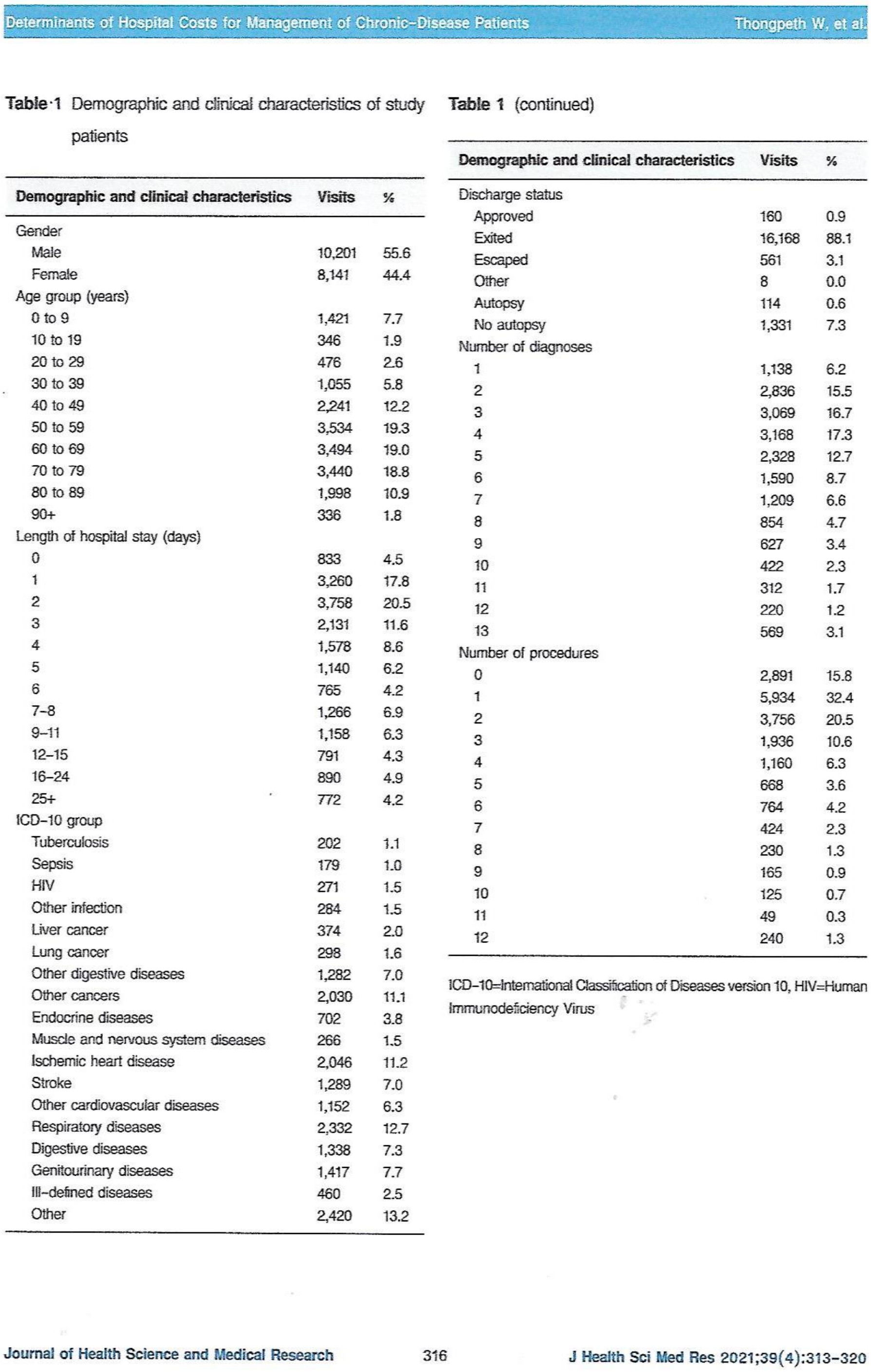


Table 1 Demographic and clinical characteristics of study patients

Demographic and clinical characteristics	Visits	%
Gender		
Male	10,201	55.6
Female	8,141	44.4
Age group (years)		
0 to 9	1,421	7.7
10 to 19	346	1.9
20 to 29	476	2.6
30 to 39	1,055	5.8
40 to 49	2,241	12.2
50 to 59	3,534	19.3
60 to 69	3,494	19.0
70 to 79	3,440	18.8
80 to 89	1,998	10.9
90+	336	1.8
Length of hospital stay (days)		
0	833	4.5
1	3,260	17.8
2	3,758	20.5
3	2,131	11.6
4	1,578	8.6
5	1,140	6.2
6	765	4.2
7–8	1,266	6.9
9–11	1,158	6.3
12–15	791	4.3
16–24	890	4.9
25+	772	4.2
ICD-10 group		
Tuberculosis	202	1.1
Sepsis	179	1.0
HIV	271	1.5
Other infection	284	1.5
Liver cancer	374	2.0
Lung cancer	298	1.6
Other digestive diseases	1,282	7.0
Other cancers	2,030	11.1
Endocrine diseases	702	3.8
Muscle and nervous system diseases	266	1.5
Ischemic heart disease	2,046	11.2
Stroke	1,289	7.0
Other cardiovascular diseases	1,152	6.3
Respiratory diseases	2,332	12.7
Digestive diseases	1,338	7.3
Genitourinary diseases	1,417	7.7
Ill-defined diseases	460	2.5
Other	2,420	13.2

Table 1 (continued)

Demographic and clinical characteristics	Visits	%
Discharge status		
Approved	160	0.9
Exited	16,168	88.1
Escaped	561	3.1
Other	8	0.0
Autopsy	114	0.6
No autopsy	1,331	7.3
Number of diagnoses		
1	1,138	6.2
2	2,836	15.5
3	3,069	16.7
4	3,168	17.3
5	2,328	12.7
6	1,590	8.7
7	1,209	6.6
8	854	4.7
9	627	3.4
10	422	2.3
11	312	1.7
12	220	1.2
13	569	3.1
Number of procedures		
0	2,891	15.8
1	5,934	32.4
2	3,756	20.5
3	1,936	10.6
4	1,160	6.3
5	668	3.6
6	764	4.2
7	424	2.3
8	230	1.3
9	165	0.9
10	125	0.7
11	49	0.3
12	240	1.3

ICD-10=International Classification of Diseases version 10, HIV=Human Immunodeficiency Virus

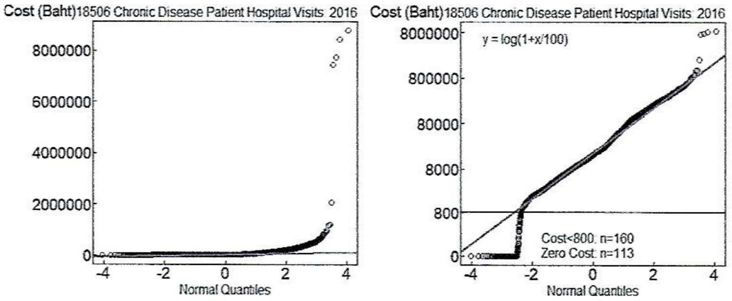


Figure 1 Normal quantile–quantile plots of cost (left) and transformed cost (right)

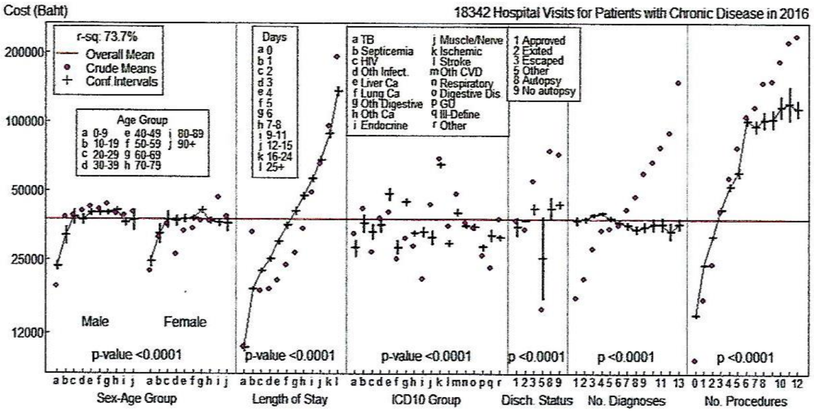


Figure 2 95% confidence interval plot of medical costs and determinants from a multiple linear regression model

As indicated above, abnormally low costs below 800 baht and abnormally high costs higher than 7 million were excluded from further analysis and a log–linear model was then created. Gender–age group, LOS, ICD–10 group, discharge status, number of diagnoses (nDiag) and number of procedures (nProc) were included in the model as determinants. Figure 2 shows the results from the multiple linear regression model.

A model was constructed to predict the natural log of the medical cost of each variable. The CI plots showed predictive accuracy of 73.7%, and the best predictors were nProc and LOS. Diagnosis (ICD–10 group) and nDiag predicted poorly when nProc and LOS were included. Crude means (circle dots) for nDiag suggested that it was a good predictor of cost. However, this correlation disappeared when LOS and nProc were included in the model. Therefore, it was found to be a confounding variable. Simple linear models were then constructed for each determinant and the r–squared values showed that nProc had the highest predictive value, of 54.1%, followed by LOS with 43.0% and nDiag with 17.6%. Since nProc and LOS had the highest r–squared values they were the only variables included in the final model even though significant p–values were found for all the other variables. nProc was classified into six groups: 0, 1, 2, 3, 4–5 and 6–12 procedures and was combined with LOS to produce a new variable named nProc–LOS with 72 groups. A simple linear regression was created with log of cost as the outcome variable and nProc–LOS as the determinant. A 95 % CI plot was created to show the relationship between cost and nProc–LOS as shown in Figure 3, which shows the CIs for hospital visit costs for 72 combinations of LOS and nProc.

Even though only two of the factors from the six–factor model were included in the final model, the r–squared only decreased by 0.015. The results therefore showed that medical costs increased when the LOS increased for all combinations of nProc and LOS except for those patients who experienced between 6 and 12 procedures during their hospitalization.

Discussion

This study explored the factors associated with hospitalization costs among chronic–disease patients using hospital administrative data. A log–linear model to estimate costs based on gender–age, diagnosis, nDiag, discharge status, LOS and nProc fit the data well with a predictive accuracy of 73.7% and all of these predictors were significantly associated with the cost, with the highest predictive value for nProc (54.1%) and LOS (43.0%) found in simple linear regressions. A reduced model with just one predictor – a factor combining nProc and LOS – produced a predictive accuracy of 72.2% with only a reduction in r–squared of 0.015. The results from the final model therefore showed that medical costs increased when the LOS and nProc increased except for those patients who experienced more than 6–12 procedures.

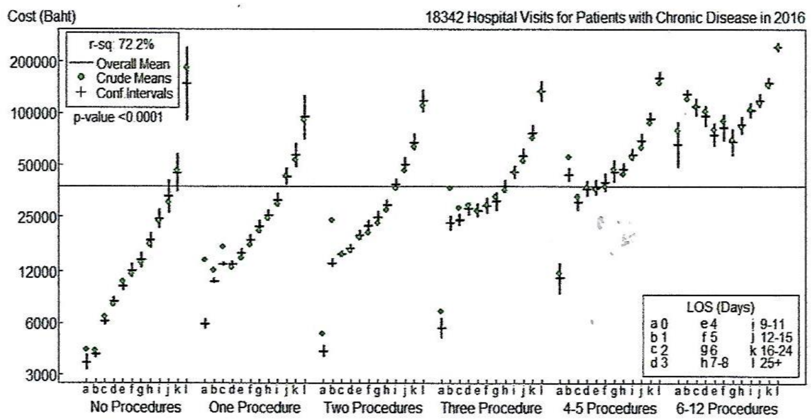


Figure 3 95% confidence interval plot of the relationship between medical cost and nProc–LOS

In this study, the patient demographic factors (gender and age), diagnosis history (principal diagnosis, and complications and comorbidities) were significantly associated with medical care costs and this result is consistent with a previous study of factors affecting health-care costs and hospitalizations among diabetic patients in Thai public hospitals conducted by Chaikledkaew et al.[10] Similar results were also reported by Slabaugh et al.[11] which found that clinical and demographic characteristics were the strong predictors of health-care cost among type 2 diabetes patients in the United States. However, the results from the present study show that the difference in the r-squared value in a model including these variables and one from which these determinants were excluded was only 0.015. Thus, it is not recommended that consideration should be given to all determinants with significant p-values, but that the overall effect on the r-squared value should be the main factor taken into consideration in determining the predictors of health-care costs since when dealing with large samples significant results can be found even where the r-squared value is quite low.

The final results showed that the determinants of medical costs among chronic-disease patients with the highest levels of significance were nProc and LOS. Thus, in assessing medical costs in Thailand, nProc and LOS should be the main factors employed in calculating the actual costs for patients. DRG-payment assessments which rely on coding systems based on diagnoses and procedures may therefore not represent accurate means of assessing patient costs when those patients are suffering from chronic diseases. In many countries where poorly developed hospital cost-accounting systems produce only low quality data, DRG systems based on those used in the United State are applied, even though they may not reflect their own practice patterns.[4] In Thailand, Pongpirul et al.[12] suggested that high quality DRG codes should not be presumed especially in resource-limited hospitals.

Conclusion

The results of this study suggest that DRG cost-assessment systems, in which costs are assessed based on the patient's diagnosis, discharge status and gender and age might not be the best means of assessing medical costs for patients with chronic illnesses in Thailand, and that the period spent in hospital and the number of procedures carried out during that time are more accurate indicators of the true medical cost.

Acknowledgement

This study was partially funded by the Graduate School, Prince of Songkla University, Thailand. We would like to give particular thanks to Suratthani Hospital, Ministry of Public Health. We are indebted to Don McNeil for giving invaluable advice regarding this research. We would like to thank the Research and Development Office, Prince of Songkla University for providing an English proofreading service.

Conflict of interest

This study has no conflicts of interest.

Funding sources

The Graduate School, Prince of Songkla University, Thailand.

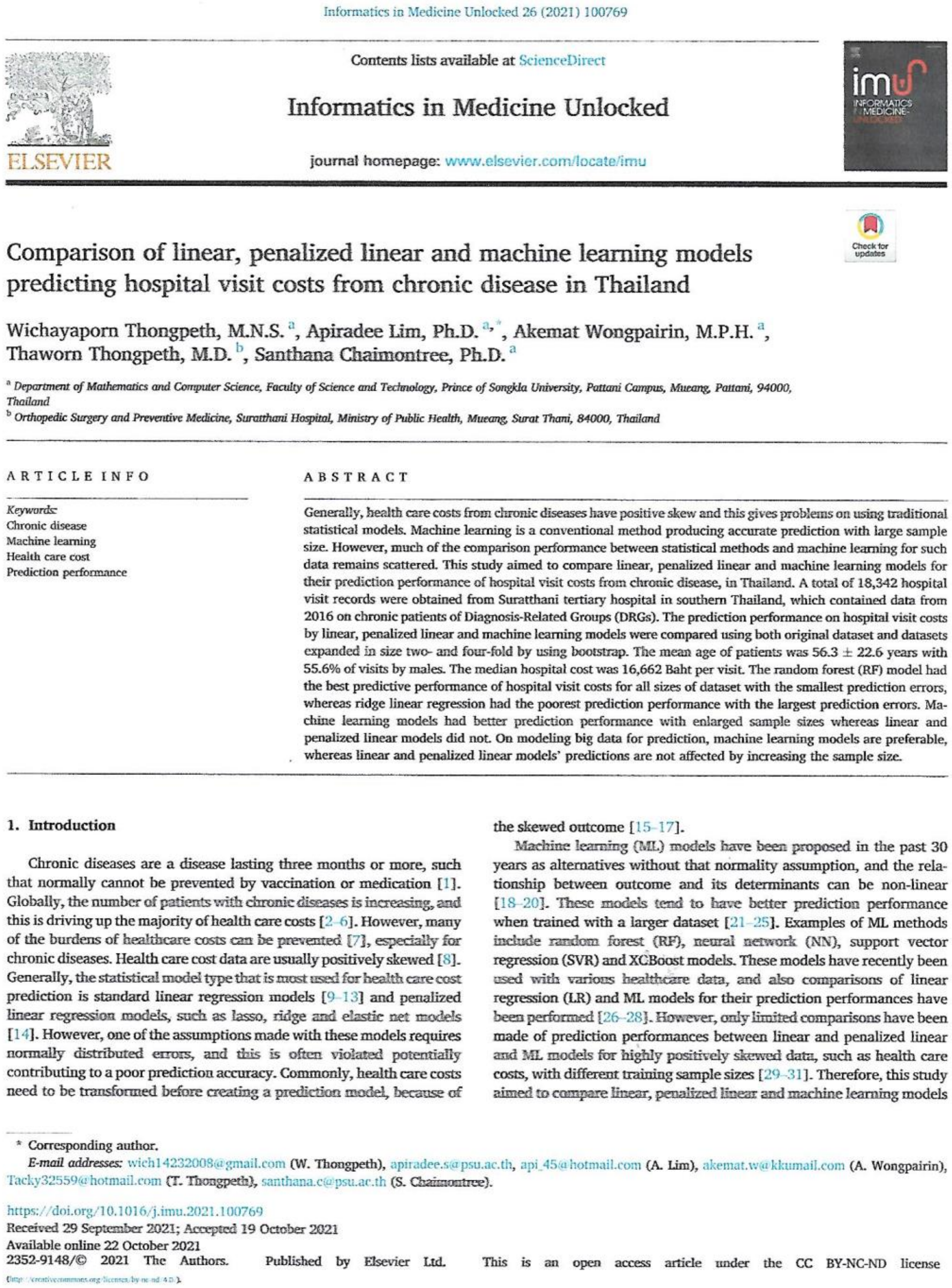
References

1. Mihailovic N, Kocic S, Jakovljevic M. Review of diagnosis-related group-based financing of hospital care. Health Serv Res Managerial Epidemiol 2016;3:1–8.
2. Chilingerian J. Origins of DRGs in the United States: a technical, political and cultural story. In: Kimberly J, de Pouvourville G, D'Aunno T, editors. The globalization of managerial innovation in health care. Cambridge: Cambridge University Press; 2008.
3. Hughes JS, Lichtenstein J, Fetter RB. Procedure codes: potential modifiers of diagnosis-related groups. Health Care Financ Rev 1990;12:39–46.

4. Scheller-Kreinsen D, Geissler A, Busse R. The ABC of DRGs. Euro Observer 2009;11:1–5.
5. Schreyögg J, Stargardt T, Tiemann O, Busse R. Methods to determine reimbursement rates for diagnosis related groups (DRG): a comparison of nine European countries. Health Care Manag Sci 2006;9:215–23.
6. Mathauer I, Wittenbecher F. Hospital payment systems based on diagnosis-related groups: experiences in low- and middle-income countries. Bull World Health Organ 2013;91:746–56A.
7. Allen C. Health education: a quick reference the go-to book for teachers. 2nd ed. Racine: Lulu Press; 2017.
8. Sakunphanit T. Universal health care coverage through pluralistic approaches: experience from Thailand. Bangkok: National Health Security Office; 2015.
9. R Development Core Team. A language and environment for statistical computing version 3.1.3. Vienna: R Foundation for Statistical Computing; 2015.
10. Chaikledkaew U, Pongchareonsuk P, Chaiyakunapruk N, Ongphiphadhanakul B. Factors affecting health-care costs and hospitalizations among diabetic patients in Thai public hospitals. Value Health 2008;11:S69–74.
11. Slabaugh LS, Curtis BH, Clore G, Fu H, Schuster DP. Factors associated with increased healthcare costs in Medicare Advantage patients with type 2 diabetes enrolled in a large representative health insurance plan in the US. J Med Econs 2015;18:106–12.
12. Pongpirul K, Walker DG, Rahman H, Robinson C. DRG coding practice: a nationwide hospital survey in Thailand. BMC Health Serv Res 2011;11:290.

Appendix II Article

Comparison of Linear, Penalized Linear and Machine Learning Models Predicting Hospital Visit Costs from Chronic Disease in Thailand”



Informatics in Medicine Unlocked 26 (2021) 100769

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu

Comparison of linear, penalized linear and machine learning models predicting hospital visit costs from chronic disease in Thailand

Wichayaporn Thongpeth, M.N.S. [a], Apiradee Lim, Ph.D. [a,*], Akemat Wongpairin, M.P.H. [a], Thaworn Thongpeth, M.D. [b], Santhana Chaimontree, Ph.D. [a]

[a] *Department of Mathematics and Computer Science, Faculty of Science and Technology, Prince of Songkla University, Pattani Campus, Mueang, Pattani, 94000, Thailand*

[b] *Orthopedic Surgery and Preventive Medicine, Suratthani Hospital, Ministry of Public Health, Mueang, Surat Thani, 84000, Thailand*

ARTICLE INFO

Keywords:
Chronic disease
Machine learning
Health care cost
Prediction performance

ABSTRACT

Generally, health care costs from chronic diseases have positive skew and this gives problems on using traditional statistical models. Machine learning is a conventional method producing accurate prediction with large sample size. However, much of the comparison performance between statistical methods and machine learning for such data remains scattered. This study aimed to compare linear, penalized linear and machine learning models for their prediction performance of hospital visit costs from chronic disease, in Thailand. A total of 18,342 hospital visit records were obtained from Suratthani tertiary hospital in southern Thailand, which contained data from 2016 on chronic patients of Diagnosis-Related Groups (DRGs). The prediction performance on hospital visit costs by linear, penalized linear and machine learning models were compared using both original dataset and datasets expanded in size two- and four-fold by using bootstrap. The mean age of patients was 56.3 ± 22.6 years with 55.6% of visits by males. The median hospital cost was 16,662 Baht per visit. The random forest (RF) model had the best predictive performance of hospital visit costs for all sizes of dataset with the smallest prediction errors, whereas ridge linear regression had the poorest prediction performance with the largest prediction errors. Machine learning models had better prediction performance with enlarged sample sizes whereas linear and penalized linear models did not. On modeling big data for prediction, machine learning models are preferable, whereas linear and penalized linear models' predictions are not affected by increasing the sample size.

1. Introduction

Chronic diseases are a disease lasting three months or more, such that normally cannot be prevented by vaccination or medication [1]. Globally, the number of patients with chronic diseases is increasing, and this is driving up the majority of health care costs [2–6]. However, many of the burdens of healthcare costs can be prevented [7], especially for chronic diseases. Health care cost data are usually positively skewed [8]. Generally, the statistical model type that is most used for health care cost prediction is standard linear regression models [9–13] and penalized linear regression models, such as lasso, ridge and elastic net models [14]. However, one of the assumptions made with these models requires normally distributed errors, and this is often violated potentially contributing to a poor prediction accuracy. Commonly, health care costs need to be transformed before creating a prediction model, because of the skewed outcome [15–17].

Machine learning (ML) models have been proposed in the past 30 years as alternatives without that normality assumption, and the relationship between outcome and its determinants can be non-linear [18–20]. These models tend to have better prediction performance when trained with a larger dataset [21–25]. Examples of ML methods include random forest (RF), neural network (NN), support vector regression (SVR) and XGBoost models. These models have recently been used with various healthcare data, and also comparisons of linear regression (LR) and ML models for their prediction performances have been performed [26–28]. However, only limited comparisons have been made of prediction performances between linear and penalized linear and ML models for highly positively skewed data, such as health care costs, with different training sample sizes [29–31]. Therefore, this study aimed to compare linear, penalized linear and machine learning models

[*] Corresponding author.
E-mail addresses: wich14232008@gmail.com (W. Thongpeth), apiradee.s@psu.ac.th, api_45@hotmail.com (A. Lim), akemat.w@kkumail.com (A. Wongpairin), Tacky32559@hotmail.com (T. Thongpeth), santhana.c@psu.ac.th (S. Chaimontree).

https://doi.org/10.1016/j.imu.2021.100769
Received 29 September 2021; Accepted 19 October 2021
Available online 22 October 2021

2352-9148/© 2021 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

for their prediction performances of hospital visit costs from chronic diseases in Thailand.

2. Methods

2.1. Data source

Health visit cost data in the year 2016 were obtained from Suratthani tertiary hospital database in southern Thailand, on claims of health care costs per capita from the Thailand National Health Security Office. A total of 18,342 admission records from chronic diseases were included in this study. The variables in this dataset include patient's age, gender, admission date, discharge date, discharge status, number of principal diagnoses, and secondary diagnosis range from 0 to 12 diagnoses, number of treatment procedures range from 0 to 12 procedures, and the total visit cost.

Gender was classified as male or female. Age was binned into ten groups with 10-year intervals: 0–9, 10–19, 20–29, 30–39, 40–49, 50–59, 60–69, 70–79, 80–89 and 90 and older. Gender and age groups were combined to a new variable with 20 groups (2 × 10) called gender-age group. ICD-10 was classified into 18 groups as mentioned in a previous study [32] and was ordered from the highest to the lowest frequency where the last group combined all other small frequencies together. This variable is called disease-group. Discharge status fell into six groups: approved by physicians, denied treatment, escaped from the hospital, died with autopsy, died without autopsy, and other. The total cost in Thai Baht (1 USD is about 31 THB) per patient-visit is the outcome. The determinants are gender-age group, length of hospital stay, disease-group, discharge status, number of diagnoses, and number of procedures. All of these candidate predictors were selected for evaluating hospital cost in the DRG system by fitting predictive models.

Data preparation processes included data cleaning, data integration, and data transformations.

2.2. Statistical and ML models

There were 8 model types tested in this study. LR is a statistical model used as a benchmark compared to 3 penalized linear and 4 ML model types: SVR, NN, RF and XGBoost. Hospital visit costs were transformed using natural logarithm, after adding one to avoid overflow from taking the logarithm of zero costs. Gender-age group, LOS, disease-group, discharge status, number of diagnoses and number of procedures were used as determinants (i.e. predictor variables) for all model types. These models were applied to 3 sizes of dataset: original; data doubled in size by bootstrap; and similarly four-fold expanded dataset. Bootstrap technique is a method of resampling a dataset with replacement. This study applied bootstrap to increase sample sizes as based on the law of large numbers, resampling large enough dataset will approximate well the population parameters [33]. All of these datasets were randomly split into training and testing datasets with a 70:30 ratio. In this study, we compared standard LR, a statistical parametric model with three types of penalized LR: lasso, ridge and elastic net, and four types of ML models: SVR, NN, RF and XGBoost.

1) Linear regression

LR is a traditional statistical model used to model the relationship between a continuous outcome and determinants that can be continuous or categorical variables. It is a parametric model assuming a linear relationship between the outcome and the determinants, and that the errors are normally distributed and have constant variance. The model takes the following form

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \tag{1}$$

where y_i is the continuous outcome value of subject i, β_0 is intercept, β_j is the coefficient of determinant j and x_{ij} is determinant j of subject i. The unknown β_j can be estimated by minimizing the residual sum of squares as follows.

$$\widehat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \tag{2}$$

2) Lasso regression

Lasso stands for Least Absolute Shrinkage and Selection Operator. Lasso uses shrinkage as the regularization technique to reduce overfitting and complexity of the model, by reducing the coefficients of less contributive variables towards or to zero. Lasso regression is a type of linear regression and is also called penalized regression with L1 regularization [34]. The penalty term used in this model is sum of the absolute weights. The lasso unknown β_j can be estimated minimizing this objective function:

$$\widehat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}, \lambda \geq 0 \tag{3}$$

Here λ is the amount of shrinkage. If λ equals zero we recover the standard linear regression.

3) Ridge regression

Ridge regression is used for reducing the complexity of the model with L2 regularization. Ridge regression is similar to lasso regression, the only difference is the penalty term used in this model namely the sum of squared weights [34]. The ridge unknown β_j can be estimated by minimizing this objective function:

$$\widehat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}, \lambda \geq 0. \tag{4}$$

4) Elastic net regression

Elastic net regression combines lasso and ridge regression methods together resulted and handle bias better than lasso or ridge regressions [35]. The elastic net unknowns β_j can be estimated by minimizing this objective function:

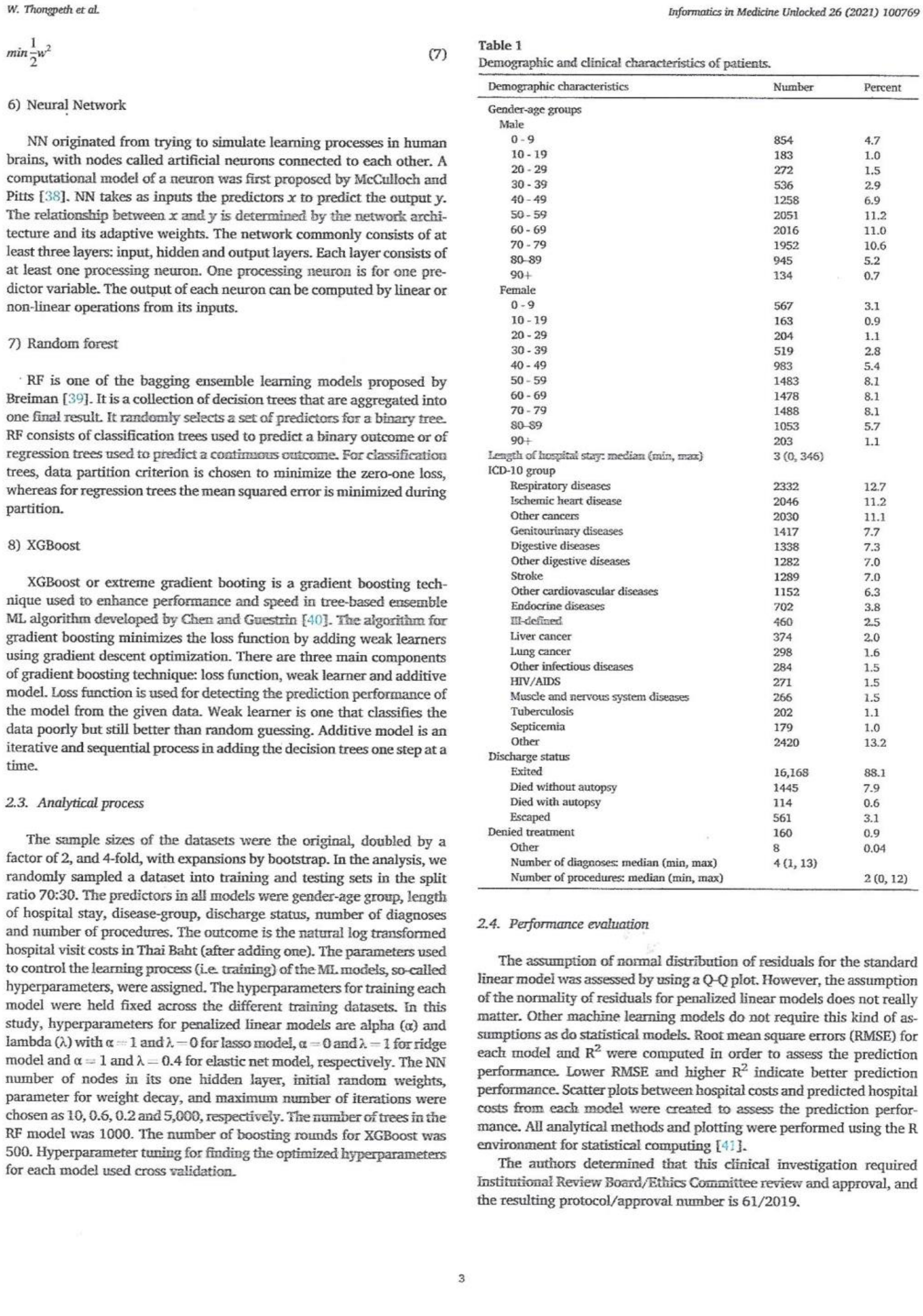
$$\widehat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum x_{ij}\beta_j \right)^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \right\}. \tag{5}$$

5) Support vector regression

SVR has similar principles to support vector machine which is used for binary outcomes. SVR is a supervised learning model proposed by Vapnik [36] used for regression problems in ML. SVR creates a set of hyperplanes in a high-dimensional space using a non-linear transformation based on the following function [37].

$$f(x) = w.x + b \tag{6}$$

Here $x \in X$ is a vector of the input predictors, $w \in X$ is the weight vector of x, and b is the error which determines the distance of the hyperplane from the original. SVR is based on minimizing the prediction error by minimizing the distance between the predicted and given output values. Thus, it uses b as the constraint to control the magnitude of the normal weight vector as follows.



$$min\frac{1}{2}w^2 \tag{7}$$

6) Neural Network

NN originated from trying to simulate learning processes in human brains, with nodes called artificial neurons connected to each other. A computational model of a neuron was first proposed by McCulloch and Pitts [38]. NN takes as inputs the predictors x to predict the output y. The relationship between x and y is determined by the network architecture and its adaptive weights. The network commonly consists of at least three layers: input, hidden and output layers. Each layer consists of at least one processing neuron. One processing neuron is for one predictor variable. The output of each neuron can be computed by linear or non-linear operations from its inputs.

7) Random forest

RF is one of the bagging ensemble learning models proposed by Breiman [39]. It is a collection of decision trees that are aggregated into one final result. It randomly selects a set of predictors for a binary tree. RF consists of classification trees used to predict a binary outcome or of regression trees used to predict a continuous outcome. For classification trees, data partition criterion is chosen to minimize the zero-one loss, whereas for regression trees the mean squared error is minimized during partition.

8) XGBoost

XGBoost or extreme gradient booting is a gradient boosting technique used to enhance performance and speed in tree-based ensemble ML algorithm developed by Chen and Guestrin [40]. The algorithm for gradient boosting minimizes the loss function by adding weak learners using gradient descent optimization. There are three main components of gradient boosting technique: loss function, weak learner and additive model. Loss function is used for detecting the prediction performance of the model from the given data. Weak learner is one that classifies the data poorly but still better than random guessing. Additive model is an iterative and sequential process in adding the decision trees one step at a time.

2.3. Analytical process

The sample sizes of the datasets were the original, doubled by a factor of 2, and 4-fold, with expansions by bootstrap. In the analysis, we randomly sampled a dataset into training and testing sets in the split ratio 70:30. The predictors in all models were gender-age group, length of hospital stay, disease-group, discharge status, number of diagnoses and number of procedures. The outcome is the natural log transformed hospital visit costs in Thai Baht (after adding one). The parameters used to control the learning process (i.e. training) of the ML models, so-called hyperparameters, were assigned. The hyperparameters for training each model were held fixed across the different training datasets. In this study, hyperparameters for penalized linear models are alpha (α) and lambda (λ) with $\alpha = 1$ and $\lambda = 0$ for lasso model, $\alpha = 0$ and $\lambda = 1$ for ridge model and $\alpha = 1$ and $\lambda = 0.4$ for elastic net model, respectively. The NN number of nodes in its one hidden layer, initial random weights, parameter for weight decay, and maximum number of iterations were chosen as 10, 0.6, 0.2 and 5,000, respectively. The number of trees in the RF model was 1000. The number of boosting rounds for XGBoost was 500. Hyperparameter tuning for finding the optimized hyperparameters for each model used cross validation.

Table 1
Demographic and clinical characteristics of patients.

Demographic characteristics	Number	Percent
Gender-age groups		
Male		
0 - 9	854	4.7
10 - 19	183	1.0
20 - 29	272	1.5
30 - 39	536	2.9
40 - 49	1258	6.9
50 - 59	2051	11.2
60 - 69	2016	11.0
70 - 79	1952	10.6
80-89	945	5.2
90+	134	0.7
Female		
0 - 9	567	3.1
10 - 19	163	0.9
20 - 29	204	1.1
30 - 39	519	2.8
40 - 49	983	5.4
50 - 59	1483	8.1
60 - 69	1478	8.1
70 - 79	1488	8.1
80-89	1053	5.7
90+	203	1.1
Length of hospital stay: median (min, max)	3 (0, 346)	
ICD-10 group		
Respiratory diseases	2332	12.7
Ischemic heart disease	2046	11.2
Other cancers	2030	11.1
Genitourinary diseases	1417	7.7
Digestive diseases	1338	7.3
Other digestive diseases	1282	7.0
Stroke	1289	7.0
Other cardiovascular diseases	1152	6.3
Endocrine diseases	702	3.8
Ill-defined	460	2.5
Liver cancer	374	2.0
Lung cancer	298	1.6
Other infectious diseases	284	1.5
HIV/AIDS	271	1.5
Muscle and nervous system diseases	266	1.5
Tuberculosis	202	1.1
Septicemia	179	1.0
Other	2420	13.2
Discharge status		
Exited	16,168	88.1
Died without autopsy	1445	7.9
Died with autopsy	114	0.6
Escaped	561	3.1
Denied treatment	160	0.9
Other	8	0.04
Number of diagnoses: median (min, max)	4 (1, 13)	
Number of procedures: median (min, max)		2 (0, 12)

2.4. Performance evaluation

The assumption of normal distribution of residuals for the standard linear model was assessed by using a Q-Q plot. However, the assumption of the normality of residuals for penalized linear models does not really matter. Other machine learning models do not require this kind of assumptions as do statistical models. Root mean square errors (RMSE) for each model and R^2 were computed in order to assess the prediction performance. Lower RMSE and higher R^2 indicate better prediction performance. Scatter plots between hospital costs and predicted hospital costs from each model were created to assess the prediction performance. All analytical methods and plotting were performed using the R environment for statistical computing [41].

The authors determined that this clinical investigation required Institutional Review Board/Ethics Committee review and approval, and the resulting protocol/approval number is 61/2019.

Table 2
RMSE and R^2 for statistical and ML models applied to different size datasets.

Model	Training		Testing	
	RMSE	R^2	RMSE	R^2
Original data				
LR	0.5945	74.1	0.5914	73.0
Lasso	0.5892	74.1	0.6041	73.0
Ridge	0.6035	73.1	0.6185	71.8
Elastic net	0.5892	73.1	0.6042	73.0
SVR	0.5969	74.2	0.5957	72.8
NN	0.4818	83.1	0.5374	78.4
RF	0.4151	87.8	0.5178	79.5
XGBoost	0.4056	87.8	0.5291	79.3
Two times expansion				
LR	0.5858	73.5	0.5872	73.1
Lasso	0.5945	73.5	0.6030	73.1
Ridge	0.6088	72.4	0.6150	72.1
Elastic net	0.5945	72.4	0.6030	73.1
SVR	0.5604	75.7	0.5637	75.1
NN	0.4688	82.9	0.5033	80.2
RF	0.3418	91.4	0.4093	87.2
XGBoost	0.4006	88.0	0.4526	84.8
Four times expansion				
LR	0.5919	73.8	0.5880	73.4
Lasso	0.5941	73.8	0.5955	73.4
Ridge	0.6081	72.6	0.6092	72.3
Elastic net	0.6082	72.8	0.6092	73.4
SVR	0.5507	77.3	0.5500	76.8
NN	0.4784	82.7	0.4922	81.5
RF	0.3125	92.8	0.3542	90.6
XGBoost	0.4115	87.5	0.4350	85.8

RMSE = Root mean square error, R^2 = R-squared, LR = Linear regression. SVR = Support vector regression, NN = Neural network, RF = Random forest.

3. Results

About 55.6% of patients were males. Most of them were aged between 50 and 79 years, accounting for 57.0%. The median length of hospital stay (LOS) was 3 days. Respiratory diseases, ischemic heart disease and cancers other than liver or lung cancer were found for 12.7%, 11.2% and 11.1%, respectively. The median number of procedure was 2 (min = 0, max = 12) while the median number of diagnosis was 4 (min = 1, max = 13) (see Table 1).

The prediction performance of each model for all sample sizes is shown in Table 2. The results for the original dataset show that RF and XGBoost had superior prediction performance in both training and testing datasets with almost the same R^2 for both, followed by NN as shown in Table 2, Fig. 1a and Fig. 1b. LR, lasso, elastic net and SVR provided almost similar performance in both training and testing datasets whereas ridge regression provided the lowest R^2 in testing dataset.

After doubling the sample size, RF had the best prediction performance, as shown in Fig. 2a and Fig. 2b. The results show that RF had superior prediction performance in both training and testing datasets with R^2 of 0.914 and 0.872 and RMSE of 0.3418 and 0.4093, followed by XGBoost with R^2 of 0.880 and 0.848 and RMSE of 0.4006 and 0.4526, NN with R^2 of 0.829 and 0.802 and RMSE of 0.4688 and 0.5033 and SVR with R^2 of 0.757 and 0.751 and RMSE of 0.5604 and 0.5637. Ridge regression had the poorest prediction performance in both training and testing for the doubled dataset, with R^2 of 0.724 and 0.721 and RMSE of 0.6088 and 0.6150.

After increasing the sample size four-fold, the RF still had the best prediction performance, as shown in Fig. 3a and Fig. 3b. The results show that RF was superior in both training and testing datasets with R^2 of 0.928 and 0.906 and RMSE of 0.3125 and 0.3542, followed by XGBoost with R^2 of 0.875 and 0.858 and RMSE of 0.4115 and 0.4350, and NN with R^2 of 0.827 and 0.815 and RMSE of 0.4784 and 0.4922. The prediction performance of SVR was ranked fourth with R^2 of 0.773 and 0.768 of RMSE of 0.5507 and 0.55 for training and testing in the four-fold expanded dataset. Ridge had the poorest prediction performance again in both training and testing with R^2 of 0.726 and 0.723 and RMSE of 0.6081 and 0.6092.

4. Discussion

In this study, prediction performances were compared between statistical and machine learning models. The findings revealed that RF model type performed best in all datasets of hospital visit costs, followed by XGBoost, NN and SVR. The LR, lasso and elastic net models had almost equal prediction performances with the original, when size was increased 2 or 4 fold by bootstrap. Ridge regression was the poorest with all sizes of datasets.

RF had the best prediction performance among the models tested. The results from our study support the findings by Lakshmanarao et al. [42] who applied LR, SVR, Decision Tree and RF to predicting medical costs, and found that RF had superior prediction performance. Kuo et al. [43] also found that RF model had the greatest accuracy when

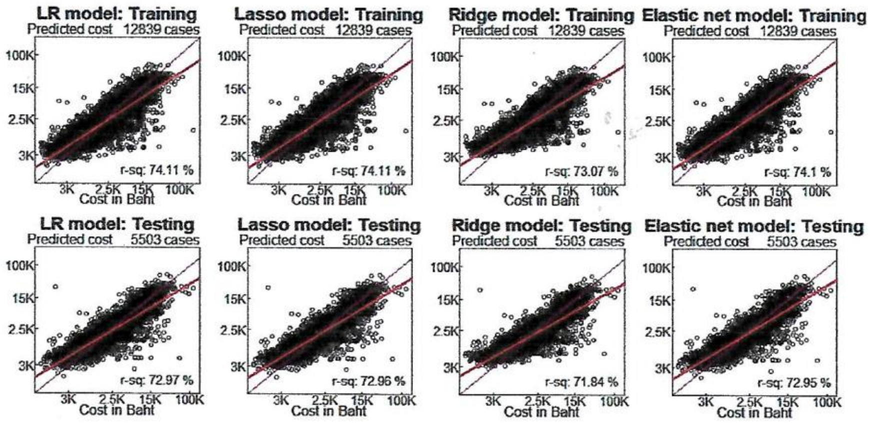


Fig. 1a. Scatter plots of predicted cost against actual cost in Thai Baht of linear and penalized linear models applied to the original dataset.

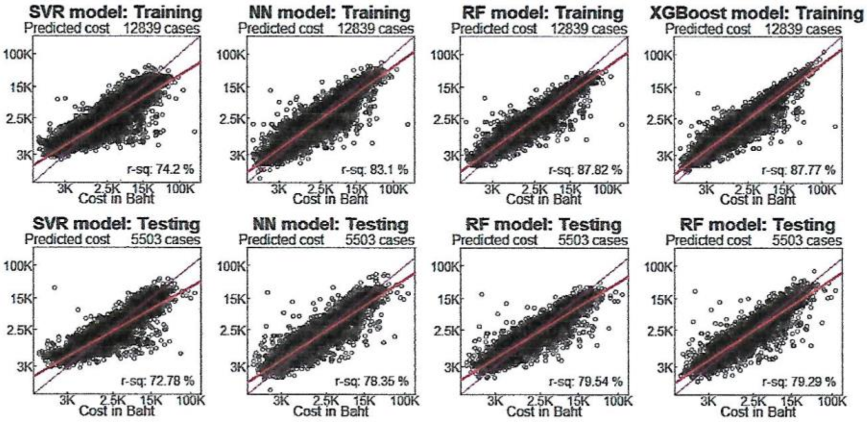


Fig. 1b. Scatter plots of predicted cost against actual cost in Thai Baht of machine learning models applied to the original dataset.

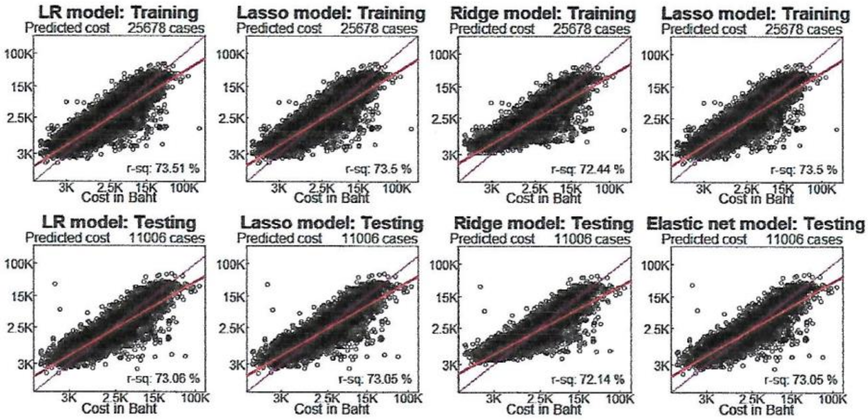


Fig. 2a. Scatter plots of predicted cost against actual cost in Thai Baht of linear and penalized linear models applied to doubled dataset.

predicting the medical costs associated with spinal fusion in terms of profit or loss in Taiwan DRGs. The prediction performance of RF depends on the strength of the individual trees in the forest, and on the correlation between them [39]. This is among the powerful ensemble ML models that tend to avoid problems with overfitting data. However, our findings are inconsistent with the results from some other studies [44,45]. Seligman et al. [45] compared LR, penalized regressions, RF and NN using social determinants of health to predict health outcomes and found that NN vastly outperformed the other three model types. This might be due to the NN being the most flexible, allowing interactions and non-linear relationships between the determinants and the outcome more than the other ML model types that are commonly used with social data.

Among the ML models, SVR provided the poorest prediction performance similar to linear, lasso and elastic net models for the original dataset. It had a better prediction performance than linear and penalized linear models when sample size was expanded by factor of 2 or 4 by bootstrap. Our findings show that linear and penalized linear models were not affected by the size of sample, as the R^2 and RMSE only slightly changed with increased sample size, whereas the expanded sample sizes had obvious effects on SVR, RF and XGBoost giving increased R^2 and decreased RMSE. However, increasing the sample size did not much affect NN. This might be because the sample size used in this study was already large enough for NN to learn. Linear and penalized linear models are methods whose performance depends on variance in the data. Even on increasing the sample size by bootstrap method, the variance in the data did not change much. Conversely, the performance of ML models depend on the size of sample. A larger sample size tends to give better prediction performance of ML models as they need a large enough sample to learn.

There are some limitations in this study. Different numbers of predictors were not considered in this study, which might also affect the

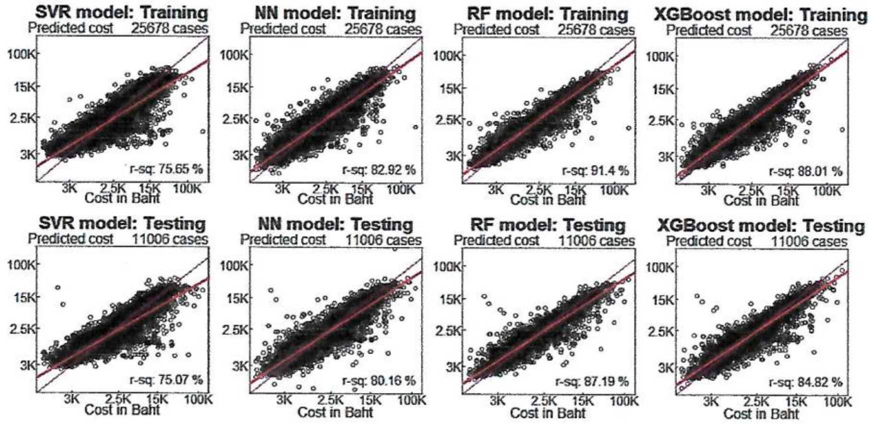


Fig. 2b. Scatter plots of predicted cost against actual cost in Thai Baht of machine learning models applied to doubled dataset.

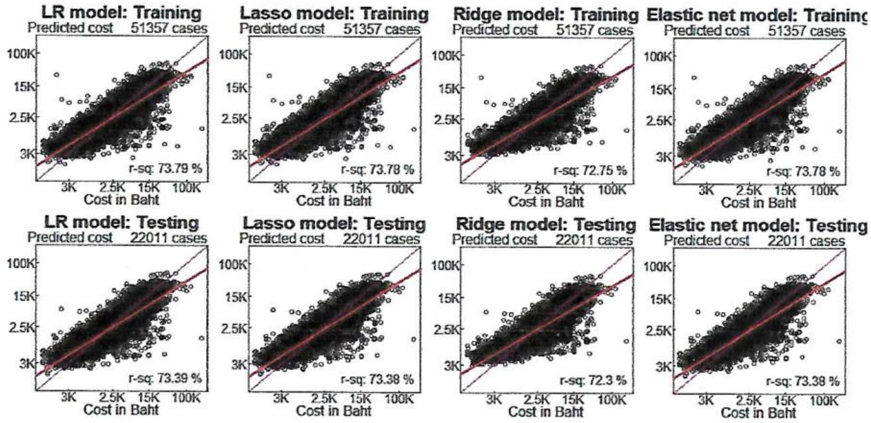


Fig. 3a. Scatter plots of predicted cost against actual cost in Thai Baht of linear and penalized linear models applied to quadruple sized dataset.

prediction performance. Lastly, the prediction performance of ML models also depends on their training hyperparameters, and this was not considered in our study. However, the hyperparameters typically have not much effect with large sized training samples.

5. Conclusions

The prediction performance of linear and penalized linear models are not affected by the increasing of training sample size, while ML models are preferred for making predictions with large training sample size.

Ethical approval

The authors determined that this clinical investigation required Institutional Review Board/Ethics Committee review and approval, and the resulting protocol/approval number was 61/2019.

Availability of data and material

The annotated MIMIC dataset is not made publicly available, because researchers are required to meet ethical conditions to access MIMIC-derived datasets. To access this dataset, please contact the corresponding author directly.

Contribution

Wichayaporn Thongpeth: Contributed to data collection, literature review, data exploration, statistical analysis, creating initial manuscript draft.

Apiradee Lim: Contributed to supervision, literature review, statistical analysis, evaluation, writing and revising the manuscript.

Akemat Wongpairin: Contributed to data exploration and literature review.

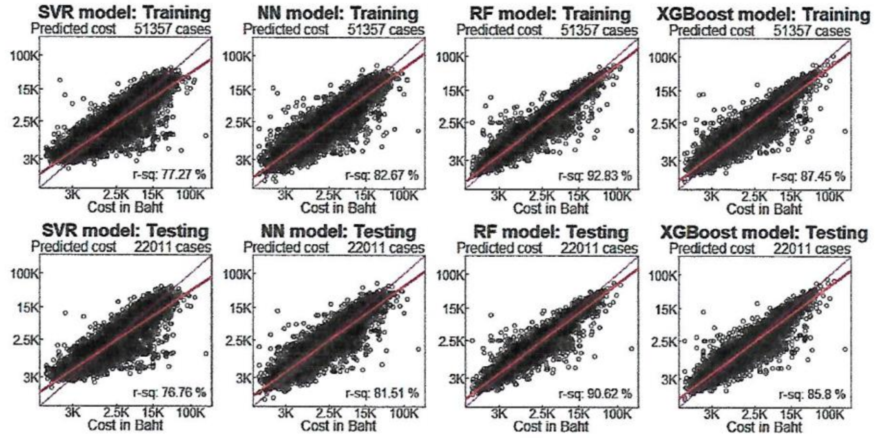


Fig. 3b. Scatter plots of predicted cost against actual cost in Thai Baht of machine learning models applied to quadruple sized dataset.

Thaworn Thongpeth: Contributed to data collection and literature review.

Santhana Chaimontree: Contributed to revising the manuscript.

Declaration of competing interest

The authors of the study declare having no conflicts of interest.

Acknowledgements

This study was funded by the Graduate School, Prince of Songkla University, Thailand. We want to give particular thanks to Suratthani Hospital, Ministry of Public Health. We are indebted to Professor Don McNeil for giving invaluable advice regarding this research. We want to thank the Research and Development Office, Prince of Songkla University, for providing a grammar checking service.

References

[1] Meetoo D. Chronic diseases: the silent global epidemic. Br J Nurs 2008;17:1320–5. https://doi.org/10.12968/bjon.2008.17.21.31731.

[2] Miranda JJ, Kinra1 S, Casas JP, Smith GD, Ebrahim S. Non-communicable diseases in low- and middle-income countries: context, determinants and health policy. T Trop Med Int Health 2008;13:1225–34. https://doi.org/10.1111/j.1365-3156.2008.02116.x.

[3] Kanken HT, Saksena P, Xu K, Evans DB. The financial burden from non-communicable diseases in low-and middle-income countries: a literature review. Health Res Pol Syst 2013;11:31. https://doi.org/10.1186/1478-4505-11-31.

[4] Slabaugh LS, Curtis BH, Clore G, Fu H, Schuster DP. Factors associated with increased healthcare costs in Medicare Advantage patients with type 2 diabetes enrolled in a large representative health insurance plan in the US. J Med Econ 2015;18:106–12. https://doi.org/10.3111/13696998.2014.979292.

[5] Chapel JM, Ritchey MD, Zhang D, Wang G. Prevalence and medical costs of chronic diseases among adult medicaid beneficiaries. Am J Prev Med 2017;53:143–54. https://doi.org/10.1016/j.amepre.2017.07.019.

[6] Holman HR. The relation of the chronic disease epidemic to the health care crisis. ACR Open Rheumatol 2019;2:1–7. https://doi.org/10.1002/acr2.11114.

[7] Glasgow RE, Orleans CT, Wagner EH. Does the chronic care model serve also as a template for improving prevention? Milbank Q 2001;79:579–612. https://doi.org/10.1111/1468-0009.00222.

[8] Dodd S, Bassi A, Bodger K, Williamson. A comparison of multivariable regression models to analyse cost data. J Eval Clin Pract 2006;12:76–86. https://doi.org/10.1111/j.1365-2753.2006.00610.x.

[9] Gertman PM, Lowenstein S. A research paradigm for severity of illness: issues for the diagnosis- related group system. Health Care Financ Rev 1984;12:79-90. PMID: 10311079; PMCID: PMC4195107.

[10] Penberthy L, Retchin SM, McDonald MK, McClish DK, Desch CE, Riley GF, Smith TJ, Hillner BE. Predictors of medicare costs in elderly beneficiaries with breast, colorectal, lung, or prostate cancer. Health Care Manag Sci 1999;2:149–60. https://doi.org/10.1023/a:1019096030306.

[11] Thorpe KE, Philyaw M. The medicalization of chronic disease and costs. Annu Rev Publ Health 2012;33:409–23. https://doi.org/10.1146/annurev-publhealth-031811-124652.

[12] Smith WM, Friedman B, Karaca Z, Wong SH. Predicting inpatient hospital payments in the United States: a retrospective analysis. BMC Health Serv Res 2015; 15:372. https://doi.org/10.1186/s12913-015-1040-8.

[13] Takeshima T, Keino S, Aoki R, Matsui T, Iwasaki K. Development of medical cost prediction model based on statistical machine learning using health insurance claims data. Value Health 2018;21:S97. https://doi.org/10.1016/j.jval.2018.07.738.

[14] Kan HJ, Kharrazi H, Chang HY, Bodycombe D, Lemke K, Weiner JP. Exploring the use of machine learning for risk adjustment: a comparison of standard and penalized linear regression models in predicting health care costs in older adults. PLoS One 2019;14(3):e0213258. https://doi.org/10.1371/journal.pone.0213258.

[15] Gregori D, Petrinco M, Bo S, Desideri A, Merletti F, Pagano E. Regression models for analyzing costs and their determinants in health care: an introductory review. Int J Qual Health Care 2011;23:331–41. https://doi.org/10.1093/intqhc/mzr010.

[16] Franzco RCD, Farmer LCM. Understanding and checking the assumptions of linear regression: a primer for medical researchers. Clin Exp Ophthalmol 2014;42:590–6. https://doi.org/10.1111/ceo.12358.

[17] Malehi AS, Pourmotahari F, Angali KA. Statistical models for the analysis of skewed healthcare cost data: a simulation study. Health Econ Rev 2015;5:11. https://doi.org/10.1186/s13561-015-0045-7.

[18] Boulesteix AL, Schmid M. Machine learning versus statistical modeling. Biom J 2014;56:588–93. https://doi.org/10.1002/bimj.201300226.

[19] Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J 2015;13:8–17. https://doi.org/10.1016/j.csbj.2014.11.005.

[20] Bzdok D. Statistics versus machine learning. Nat Methods 2018;15:233–4. https://doi.org/10.1038/nmeth-4642.

[21] Povak NA, Hessburg PF, McDonnell TC, Reynolds KM, Sullivan TJ, Salter RB, Cosby BJ. Machine learning and linear regression models to predict catchment-level base cation weathering rates across the southern Appalachian Mountain region, USA. Am Geo Union 2014;50:2798–814. https://doi.org/10.1002/2013WR014203.

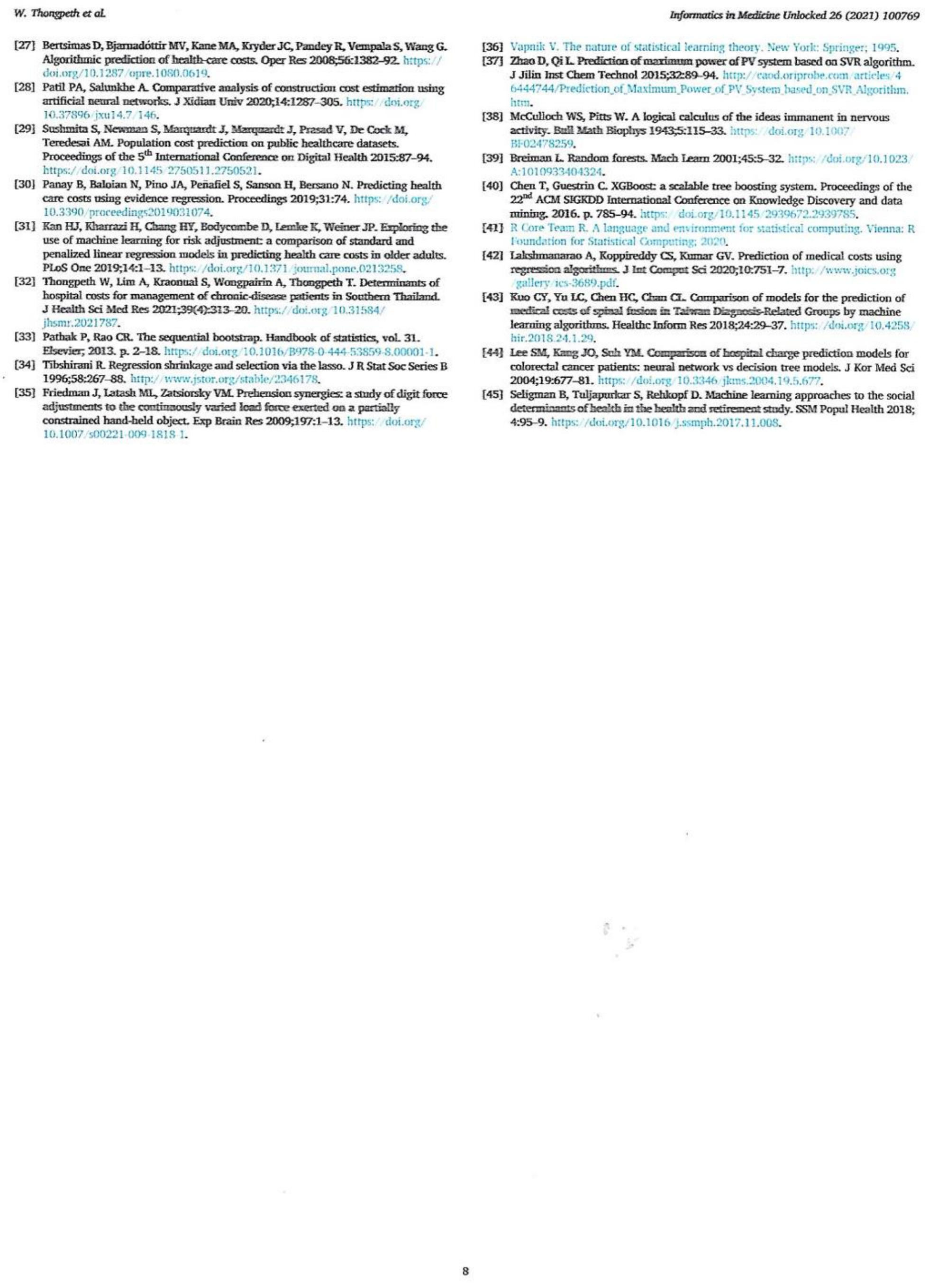
[22] Sweety BE, Srimathi H, Bagavandas M. A Survey of machine learning algorithms in health care. Int J Sci Technol Res 2019;8:2288–91. http://www.ijstr.org/final-print/nov2019/A-Survey-Of-Machine-Learning-Algorithms-In-Health-Care.pdf.

[23] Dureh N, Tongkumchum P. A comparison of logistic regression and machine learning algorithms applied to zero counts data in contingency tables. Adv Appl Stat 2019;55:67–76. https://doi.org/10.17654/AS055010067.

[24] Lim A, Taufik MR, Tongkumchum P, Dureh N. Comparison of different supervised machine learning algorithms for the prediction of tuberculosis mortality. Adv Appl Stat 2020;52:185–201. https://doi.org/10.17654/AS062020185.

[25] Rajula HSR, Verlato G, Manchia M, Antonucci N, Fanos V. Comparison of conventional statistical methods with machine learning in medicine: diagnosis, drug development, and treatment. Medicina 2020;56:455. https://doi.org/10.3390/medicina56090455.

[26] Austin PC, Ghali WA, Tu JV. A comparison of several regression models for analyzing cost of CABG surgery. Stat Med 2003;22:2799–815. https://doi.org/10.1002/sim.1442.



[27] Bertsimas D, Bjarnadóttir MV, Kane MA, Kryder JC, Pandey R, Vempala S, Wang G. Algorithmic prediction of health-care costs. Oper Res 2008;56:1382–92. https://doi.org/10.1287/opre.1080.0619.

[28] Patil PA, Salunkhe A. Comparative analysis of construction cost estimation using artificial neural networks. J Xidian Univ 2020;14:1287–305. https://doi.org/10.37896/jxu14.7/146.

[29] Sushmita S, Newman S, Marquardt J, Marquardt J, Prasad V, De Cock M, Teredesai AM. Population cost prediction on public healthcare datasets. Proceedings of the 5th International Conference on Digital Health 2015:87–94. https://doi.org/10.1145/2750511.2750521.

[30] Panay B, Baloian N, Pino JA, Peñafiel S, Sanson H, Bersano N. Predicting health care costs using evidence regression. Proceedings 2019;31:74. https://doi.org/10.3390/proceedings2019031074.

[31] Kan HJ, Kharrazi H, Chang HY, Bodycombe D, Lemke K, Weiner JP. Exploring the use of machine learning for risk adjustment: a comparison of standard and penalized linear regression models in predicting health care costs in older adults. PLoS One 2019;14:1–13. https://doi.org/10.1371/journal.pone.0213258.

[32] Thongpeth W, Lim A, Kraonual S, Wongpairin A, Thongpeth T. Determinants of hospital costs for management of chronic-disease patients in Southern Thailand. J Health Sci Med Res 2021;39(4):313–20. https://doi.org/10.31584/jhsmr.2021787.

[33] Pathak P, Rao CR. The sequential bootstrap. Handbook of statistics, vol. 31. Elsevier; 2013. p. 2–18. https://doi.org/10.1016/B978-0-444-53859-8.00001-1.

[34] Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Series B 1996;58:267–88. http://www.jstor.org/stable/2346178.

[35] Friedman J, Latash ML, Zatsiorsky VM. Prehension synergies: a study of digit force adjustments to the continuously varied load force exerted on a partially constrained hand-held object. Exp Brain Res 2009;197:1–13. https://doi.org/10.1007/s00221-009-1818-1.

[36] Vapnik V. The nature of statistical learning theory. New York: Springer; 1995.

[37] Zhao D, Qi L. Prediction of maximum power of PV system based on SVR algorithm. J Jilin Inst Chem Technol 2015;32:89–94. http://caod.oriprobe.com/articles/46444744/Prediction_of_Maximum_Power_of_PV_System_based_on_SVR_Algorithm.htm.

[38] McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. Bull Math Biophys 1943;5:115–33. https://doi.org/10.1007/BF02478259.

[39] Breiman L. Random forests. Mach Learn 2001;45:5–32. https://doi.org/10.1023/A:1010933404324.

[40] Chen T, Guestrin C. XGBoost: a scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and data mining. 2016. p. 785–94. https://doi.org/10.1145/2939672.2939785.

[41] R Core Team R. A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2020.

[42] Lakshmanarao A, Koppireddy CS, Kumar GV. Prediction of medical costs using regression algorithms. J Int Comput Sci 2020;10:751–7. http://www.joics.org/gallery/ics-3689.pdf.

[43] Kuo CY, Yu LC, Chen HC, Chan CL. Comparison of models for the prediction of medical costs of spinal fusion in Taiwan Diagnosis-Related Groups by machine learning algorithms. Healthc Inform Res 2018;24:29–37. https://doi.org/10.4258/hir.2018.24.1.29.

[44] Lee SM, Kang JO, Suh YM. Comparison of hospital charge prediction models for colorectal cancer patients: neural network vs decision tree models. J Kor Med Sci 2004;19:677–81. https://doi.org/10.3346/jkms.2004.19.5.677.

[45] Seligman B, Tuljapurkar S, Rehkopf D. Machine learning approaches to the social determinants of health in the health and retirement study. SSM Popul Health 2018;4:95–9. https://doi.org/10.1016/j.ssmph.2017.11.008.

Appendix III Proceeding

Estimating Medical Treatment Costs for Violence-related Injury in Thailand

Appendix III Proceeding

This thesis includes research results have been submitted for oral presentation in Proceedings of The ISI Regional Statistics Conference 2017 (ISI RSC 2017), 20-24 March 2017, Indonesia. The presentation title was “Estimating Medical Treatment Costs for Violence-related Injury in Thailand”. This conference sought to bring together distinguished statisticians and members of the statistical community from Southeast Asia and around the world to present, discuss, promote, and distribute research and best practices in all areas of statistics and its applications to improving human life. Provide participants with significant professional development and networking opportunities. The theme of this conference was “Health and Social Statistics”. My presentation was organized in Health and Social Statistics. Acceptance letters, the details of presented including cover and certificate are shown in this appendix.

Estimating Medical Treatment Costs for Violence-related Injury in Thailand

Wichayaporn Thongpeth,
PhD Candidate, Program in Research Methodology, Prince of Songkla University, Muang, Pattani, 94000 Thailand
wich14232008@gmail.com

Emml-Benjamin Atta Owusu Mintah,
Master degree, Program in Research Methodology, Prince of Songkla University Pattani, 94000 Thailand

Apiradee Lim
Assistant Professor, Program in Research Methodology, Prince of Songkla University,
Muang, Pattani, 94000 Thailand – api_45@hotmail.com
Department of Mathematics and Computer Science, Prince of Songkla University

Abstract

Hospitals in Thailand routinely record details of treatments for patients entering the emergency room. In the three southernmost provinces, many residents have suffered injuries from the terrorist insurgency that began in 2004. The data contained information from 7,404 subjects who made up to five visits (with 9,701 visits in total) during the five-year period from 2007 to 2011, We examined hospital data for the subjects who needed ER treatment on occasions. In 2007 the Ministry of Public Health in Thailand established a violence-related injury surveillance (VIS) database to record data on victims from the insurgency in the Deep South of Thailand (Narathiwat, Yala, and Pattani provinces and the four southernmost districts of Songkla province), but data on charges are incomplete. In this study, we illustrate a method of imputing unknown or unrecorded charges for visits using length of stay and number of diagnoses, based on a sub-sample of the patients who made visits. Factors considered are the principal diagnosis of the type of injury (ICD-10 code group) grouped by severity of injury, the number of diagnoses, and the length of stay. Results show that higher charges were incurred by victims suffering abdominal and pelvic content injuries.

The aim of the study was to analyze and compare some corresponding costs in Thailand. To do this we investigated costs for injuries caused by gunshot and bombs to persons entering emergency rooms at hospitals in southern Thailand.

The conclusions suggest that the linear regression model provides a good fit for the estimate of treatment costs based on the coefficients.

Keywords: Medical Treatment Costs; Violence-related Injury Surveillance; Linear regression model
JEL classification: G00

1. INTRODUCTION

The Violence-related Injury Surveillance (VIS) System for the Southern Border Provinces Area was established in January 2007 to develop the data system that would facilitate the development of emergency medical services, determination of strategies and plans, resource allocation, control and prevention of injuries, healing, and recovery for those who were affected by violence in the Southern Border Provinces area. The target group (population under surveillance) are all individuals who were injured or deceased from intentional injury who received treatment or whose autopsy was performed at the 48 governmental hospitals in Songkla, Satul, Pattani, Yala, and Narathiwat Provinces. Estimation of the average total cost for treating trauma patients is often complicated by the fact that the survival times are censored on some study subjects and their subsequent costs are unknown. This study presents the results of analysis of the data from the surveillance system, which only includes incidents resulting from the situation of unrest in the Southern Border Provinces. Injury deaths compared to other leading causes of mortality. The deaths caused by injuries, have an immeasurable impact on the families and communities affected, whose lives are often changed irrevocably by these tragedies. Injuries and violence have been neglected from the global health agenda for many years, despite being predictable and largely preventable. Evidence from many countries shows that dramatic successes in preventing injuries and violence can be achieved through concerted efforts that involve, but is not limited to, the health sector. The international community needs to work with governments and civil society around the world to implement these proven measures and reduce the unnecessary loss of life that occurs each day as a result of injuries and violence. Injuries are a global public health problem about 5 million people die each year as a result of injuries. Other main causes of death from injuries are falls, drowning, burns, poisoning, and war (World Health Organization, 2014).

Violence related injury surveillance also has diverse effects on the economy of many developing countries. The total cost of injuries and violence in the United States was $671 billion in 2013, according to two Morbidity and Mortality Weekly Reports (MMWR) released today by the Centers for Disease Control and Prevention (CDC). The cost associated with fatal injuries was $214 billion; nonfatal injuries were $457 billion (CDC, 2015). This lost cost 1-3% of gross national product of the government for low and middle-income countries annually. There doesn't report on the total cost of emergency medical service in Thailand. This amount is more than the total aid provided to low and middle income countries for developing health systems to prevent accidents (World Health Organization, 2008). Moreover, there is a high cost estimated benefit arising from preventing such accidents. Given the extent of this burden being confronted by low and middle-income countries, there is the need to prevent at the forefront of public health initiatives (World Health Organization, 2008). This study aimed to analyze and compare some corresponding costs in Thailand. And then to do this we investigated the cost of injuries caused by gunshot and bombs to persons entering emergency rooms at a hospital in the Depth south of Thailand.

2. MATERIALS AND METHODS

2.1. Data and variables

A retrospective analysis of Violence injury surveillance in Thailand. We obtained relevant data from the Deep South Coordination Centre (DSCC) database, Faculty of Science and Technology, Prince of Songkla University, Pattani Campus, Thailand. Data management for systematic analysis of the violence injury surveillance data were checked for errors and missing records. All of the missing values were cleaned before performing data analyses. Since cost of medical treatment was not complete recorded, we estimated the cost based on the coefficients estimated from linear regression model of VIS data. The model has three categorical determinants comprising length of stay, number of diagnosis and diagnosis group. The estimated violence related injury treatment costs were obtained and filled in the database.

2.2. Statistical methods

The relationship between diagnosis group, length of stay, and number of diagnosis of violence related injuries for each determinants will be analyzed using linear regression model. The medical treatment cost of violence related injuries will be estimated using linear regression model.

Linear regression

The linear regression can be used to explore the relationships between a dependent variable and a set of independent variables. The general method of estimation that leads to the least squares function under the linear regression model (when the error is normally distributed) is called maximum likelihood. Under the assumptions of linear regression, the method of least squares yields estimators with a number of desirable statistical properties. The specific form of the linear regression model is as equation 1.

$$y_i = a + \sum_{i=1}^{n} b_i x_i + \varepsilon_i \tag{1}$$

Where, is outcome (diagnosis group, length of stay, number of diagnosis), *a* is an intercept of y, the value of y when equals zero, $_{1}, _{2}, _{3}, \ldots, _{9}$ are, mechanism of injuries and respectively are determinants, b_1, b_2, b_3,…,b_9 *are* coefficient of each variable, ε_i is the error.

In addition, the estimate medical treatment costs from linear regression model

We call *y* the independent or response variable (estimated cost of injuries); x_1, x_2, and x_3 are the independent variables (diagnosis group, length of stay, number of diagnosis). We call *a* the "true" intercept term; b_1, b_2, and b_3 are called the slope or gradient of the line. It is the increase in y corresponding to an increase of one unit in x (Bland, 2000; McNeil, 2006).

Linear regression is a statistical method widely used to model the association between a continuous outcome and a set of fixed determinants. The model expresses the outcome variable as an additive function of the determinants. We used R software (R Development Core Team, 2011) to produce all statistical results and graphs.

2.3. Analysis strategy

To remove skewness in the linear model we transformed the incidence rates by taking their logarithms, after replacing zero counts by 0.5 to ensure finiteness. We fitted models with two additive factors as determinants for statistical reasons aimed at reducing the standard errors of the estimated parameters. This model was chosen because it is arguably more appropriate for studying.

3. RESULTS

A total of 9,701 visits and 7,404 patients were available for the economic analysis. After adjusting for patient demographic and characteristics included the number of conditions present on admission, and length of stay. As the length of stay is also often right-skewed, we transformed it to log (length of stay+1)

From this regression, we then calculated a given hospital adjusted charge for the average statewide patient for that DRG, where the adjusted charges represented standardized log charge/(day+1). This gave us a singer adjusted charge per day for each event, representing the predicted charge for a patient with the same average clinical and demographic and characteristics, which we then used as the dependent variable in our second stage regression.

4. DISCUSSION

We considered 9,701 visits from 7,404 patients entering hospitals in four provinces during year 2007-2011.We used commands in R program and structured these data as a database table indexed by each patient'13 –digit Personal Identification Number and injury and hospital entry date/time with 20 fields as follows. Main outcomes are the cost (in Baht) and final outcomes are status arrive in hospital 0 or 1: died. We separated patients into groups according to the number of occasions they visited a hospital. We gained some detailed understanding of the process, we now focus on the three subjects who made five visits using the following R commands. We viewed data for the three subjects with the most visits. We saw that the first of the three subjects. When doing a study like this it is useful to get some understanding of the process generating the data before embarking on further analysis, and carefully investigate the information available from them in detail. Next step, we analyzed the costs associated with visits from patients entering the emergency room in southern Thai hospitals for treatment of terrorism- related injuries. Since many of these charges do not appear in the database, we will develop a method for imputing these missing these missing outcomes, and illustrate this method using a relatively small sample of patients with three or four visits. We fitted model to charge using number of diagnosis and length of stay for predicting charges Charge unknown Charges have skewed distributions so need to be transformed to satisfy the normality assumption in linear models, then transformed back to get fund total. There subjects had 881 visits, their lengths of stay (LS) ranged from 0 to 215 days. But there are 171 missing values for length of stay. We created the grouped length of stay varies LS1 to have 8 levels, after imputing missing values to be the median (0). and then we created grouped variable number dignosis1 to have 5 levels by combining last three levels. The charge unknown 253 visits and their charge unreported code as -1. This data have charge having a highly skewed distribution. And then we created un-skewed variable charge1 after replacing missing values by NA and transforming using logarithms. We separated data according to charges reported or omitted. And then we fit a linear model to charges reported charges. This model accounts for 50 percentage of squared variation. The normality assumption for the errors is plausible. Use model regression coefficients to impute values of charge1 for cases with unreported charge. We merged the data tables with the reporters and imputed charges. Transform to baht. We could compare the total charges for the reported and the reported plus imputed amounts.

The estimated total charge for the 881 visits by the 284 patients who had 3 or 4 visit was 13.7 million bath. Or the average 48,289 baht per visits. Our estimate of the medical cost of each violence fatality in Thailand in 2001-2011 was 26,126,325 Baht but the government paid true cost being 50,967,541baht. We calculated the number and cost of injuries by violence related injuries using the method in Thailand, between 2007 and 2011.

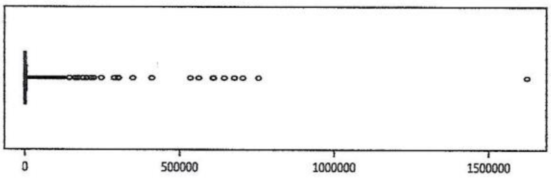


Figure 1. Box Plot Graph Show Medical Cost Among 9701 Visits (P =.005). Reported Hospital Charges from Each Visit and 75th Percentiles, with the Line Through the Box Representing the Median. The Whiskers on the Box are Equal to the 150% of the Interquartile Range Centered at the Mean

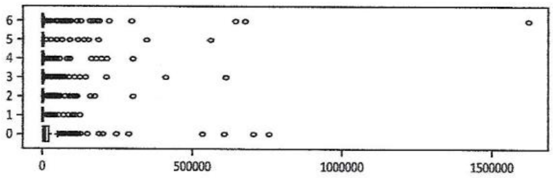


Figure 2. A Box Plot Graph Comparing Medical Cost Among 9701 Visits (P =.005). Reported Hospital Charges from Each the Number of Diagnosis. The Box is Defined by the 25th and 75th Percentiles, With the Line Through the Box RepresentingtheMedian.TheWhiskersontheBoxareEqualtothe150%oftheInterquartileRangeCenteredattheMean

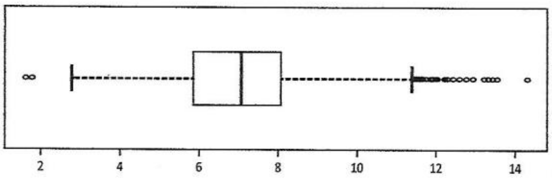


Figure 3. Box Plot Graph Comparing Mean Costs of Medical Cost Among 9701 Visits (P = .005). Reported Hospital Charges From Each the Length of Stays Were Converted to Costs by Applying Medicare Cost-to-charge Ratios. The Box is Defined by the 25th and 75th Percentiles, With the Line Through the Box Representing the Median.

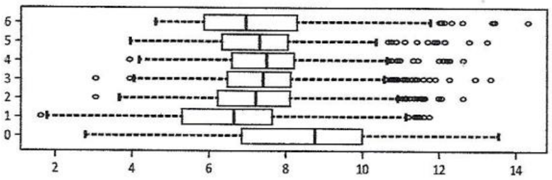


Figure 4. Box Plot Graph Comparing Mean Costs of Medical Cost Among 9701 Visits (P = .005). Reported Hospital Charges From Each the Number of Diagnosis were Converted to Costs by Applying Medicare Cost-to-charge Ratios. The Box is Defined by the 25th and 75th Percentiles, With the Line Through the Box Representing the Median. The Whiskers on the Box are Equal to the 150% of the Interquartile Range Centered at the Mean

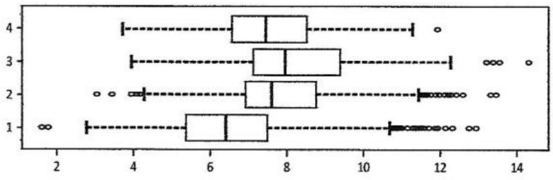


Figure 5. Box Plot Graph Comparing Mean the Length of Stay 14 Days (P = .005). Reported Hospital Charges From Each Diagnosis Group Were Converted to Costs by Applying Medicare Cost-to-charge Ratios. The Box is Defined by the 25th and 75th Percentiles, With the Line Through the Box Representing the Median. The Whiskers on the Box are Equal to the 150% of the Interquartile Range Centered at the Mean

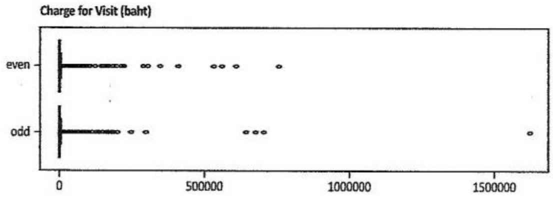


Figure 6. Box Plot Graph Comparing Mean Costs of Medical Cost Among 9701 Visits (P = .005). Reported Hospital Charges From Each Institution Were Converted to Costs by Applying Medicare Cost-to-charge Ratios.

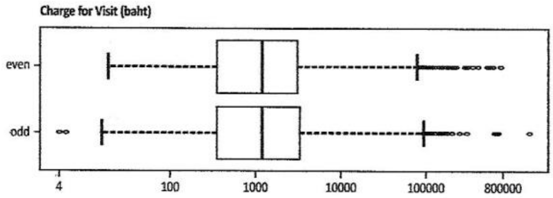


Figure 7. Box Plot Graph Comparing Mean Costs of Trauma Care Among 7404 Patients (P = .005). Reported Hospital Charges From Each Institution Were Converted to Costs by Applying Medicare Cost-to-charge Ratios. The Box is Defined by the 25th And 75th Percentiles, With the Line Through the Box Representing the Median. The Whiskers on the Box are Equal to the 150% of the Interquartile Range Centered at the Mean

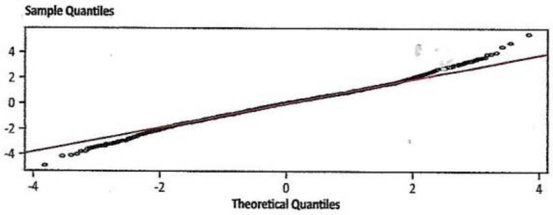


Figure 8. A Scatter Plot Showing the Data of Hospital Charge in Violence Injury Surveillance Compare the Theoretical Data.

5. CONCLUSIONS

Much of appeliez statistics may be viewed as an elaboration of the linear regression model and associated estimation methods of least-squares. In beginning to describe these techniques Mosteller and Tukey (1977) in their inuential text remark: What the regression curve does is give a grand summary for the averages of the distributions corresponding to the set of x's. We could go further and compute several different regression curves corresponding to the various percentage points of the distributions and thus get a more complete picture of the set. Ordinarily, this is not done, and so regression often gives a rather incomplete picture. Just as the mean gives an incomplete picture of a single distribution, so the regression curve gives a correspondingly incomplete picture for a set of distributionsQuantile regression methods useful to describe associations between grouped length of stay variable and hospital charge.And then it can describe associations between grouped number of diagnoses variable and hospital charge in Thailand. To do this we investigated costs for injuries caused by gunshot and bombs to persons entering emergency rooms at hospitals in southern Thailand. We suggested that the linear regression model provides a good fit for the estimate of treatment costs based on the coefficients.

REFERENCES

Ash, A.A., Ellis, R.P., Pope, G.C., Ayanian, J.Z., Bates, D.W., Burstin, H. Lezzoni, L.I., Mackay, E., and Yu, W. (2000). Using diagnosis to describe poputations and predict costs. Health Care Financing Review. 21(3), 7-28.

Brilleman, S.L., Gravelle, H., Hollinghurst, S., Purdy, S., Salisbury C., Windmeijer, F. (2014). Journal of Health Economics. 35, 109-122.

McNeil, D.(2007). Modern Statistics, Pearson SprintPrint Ltd, Australia, pp. 229-233.

Nilsen, P., Hudson, D., and Lindqvist, K.(2006). Economic analysis of injury prevention applying results and methodologies from cost-of injury studies. International Journal of Injury Control and Safety Promotion.13(1), 7–13.

World Health Organization. (2014). Special Tabulation Lists for Mortality and Morbidity; Mortality tabulation list1. In International Statistical Classification of Diseases and Health Related Problems Tenth Revision Volume 1. World Health Organization,Geneva, Switzerland, pp.2-5.

World Health Organization. (2014). Preventing Injuries and Violence: a Guide for Ministries of Health. World Health Organization, Geneva, Switzerland, pp.6-9.

World Health Organization. (2014). The Global Burden of Disease: 2004 Update. World Health Organization, Geneva, Switzerland, pp.120-121.

Yang, L., Lam, L.T., Liu, Y., Geng, W.K., and Liu, D.C. (2005). Epidemiological profile of mortality due to injuries in three cities in the Guangxi Province, in China. Accident Analysis and Prevention. 37(1), 137-141.

VITAE

Name Wichayaporn Thongpeth

Student ID 632203302

Educational Attainment

Degree	Name of Institution	Year of Graduation
B.Sc (Nursing).	Prince of Songkla University	2002
MSN. (Nursing).	Prince of Songkla University	2008

Scholarship Awards during Enrolment

- Research Grant Scholarship from Graduate School, Prince of Songkla University, Thailand

Work – Position and Address

- A Special Instructor of Faculty of Nursing, Prince of Songkla University, Pattani Campus, Pattani, Thailand.

List of Publication and Proceeding

Publication:

Thongpeth, W., Lim, A., Kraonual, S., Wongpairin, A., & Thongpeth, T. (2021). Determinants of Hospital Costs for Management of Chronic- Disease Patients in Southern Thailand. *Journal of Health Science and Medical Research,* 39(4). 313-320.

Thongpeth, W., Lim, A., Wongpairin, A., Thongpeth, T., & Chaimontree, S. (2021). Comparison of linear penalized linear and machine learning models predicting hospital visit costs from chronic disease in Thailand, *Informatics in Medicine Unlocked,* doi: 10.1016/j.imu.2021.100769.

International Seminar:

Thongpeth, W., & Lim, A. Costs for Emergency Room Patients in Southern Thailand for Patients with many Visits. International Seminar in Research Methods and Practice at The University of Malaya, Kuala Lumpur, 3-4 February 2015.

Thongpeth, W., & Lim A. Estimating the Accurate Medical Treatment Costs for Violence-related Injuries in Southern Thailand. International Research Methods in Practice Workshop and Seminar, at Prince of Songkla University Pattani Campus, 21-23 May 2017.

Thongpeth, W., & Lim, A. Determinants of Hospital Cost for Chronic Disease. International Research Methods in Practice Workshop and Seminar, at Prince of Songkla University Pattani Campus, 10-11 June 2019.

Proceeding:

Thongpeth, W., Owusu, B. A., & Lim A. Estimating Medical Treatment Costs for Violence-related Injury in Thailand. Conference on ISI Regional Statistics 20-24 March 2017 Bali, Indonesia.