Scholar Commons
USF University of South Florida

University of South Florida
Scholar Commons

Graduate Theses and Dissertations | Graduate School

January 2015

Patient Populations, Clinical Associations, and System Efficiency in Healthcare Delivery System

Yazhuo Liu
University of South Florida, yazhuoliu@mail.usf.edu

Follow this and additional works at: http://scholarcommons.usf.edu/etd

Part of the Industrial Engineering Commons, Medicine and Health Sciences Commons, and the Statistics and Probability Commons

Scholar Commons Citation
Liu, Yazhuo, "Patient Populations, Clinical Associations, and System Efficiency in Healthcare Delivery System" (2015). *Graduate Theses and Dissertations.*
http://scholarcommons.usf.edu/etd/5726

This Dissertation is brought to you for free and open access by the Graduate School at Scholar Commons. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Scholar Commons. For more information, please contact scholarcommons@usf.edu.

Patient Populations, Clinical Associations, and System Efficiency in Healthcare Delivery System

by

Yazhuo Liu

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Industrial and Management Systems Engineering
College of Engineering
University of South Florida

Co-Major Professor: Jose Zayas-Castro, Ph.D.
Co-Major Professor: Shuai Huang, Ph.D.
Peter J. Fabri, M.D.
Alex Savachkin, Ph.D.
Stephanie Carey, Ph.D.
Vic Velanovich, M.D.

Date of Approval:
June 23, 2015

Keywords: Hospital Readmissions, Nonlinear Networks, Tree Structures, Operating Room Scheduling, Stochastic Programming

Copyright © 2015, Yazhuo Liu

ACKNOWLEDGMENTS

First, I would like to thank my advisor Dr. Jose Zayas-Castro, Dr. Shuai Huang and Dr. Peter J. Fabri for providing great guidance through the past few years. This dissertation cannot be done without their help and support.

Second, I would like to acknowledge for giving me valuable advice during my PhD studies. His experiences and expertise has shed light on my research over the past few years. I also want to thank Dr. Stephanie Carey and Dr. Vic Velanovich for their support and suggestions in my dissertation.

Finally, I would like to mention my husband Long for his unconditional love, caring and hardwork; my family back in China for understanding and encouraging me over the sea; the faculty and students of the Department of Industrial and Management Systems Engineering for accepting and helping me; the University of South Florida College of Medicine, Morsani Surgery Center and Tampa General Hospital for providing me time, support and collaboration.

TABLE OF CONTENTS

LIST OF TABLES iii

LIST OF FIGURES iv

ABSTRACT v

CHAPTER 1: INTRODUCTION 1

CHAPTER 2: RISK FACTORS OF HOSPITAL READMISSIONS 4
- 2.1 Introduction 4
- 2.2 Risk Factors for Patients with Chronic Conditions 5
  - 2.2.1 Background 5
  - 2.2.2 Study Design 6
  - 2.2.3 Results 9
  - 2.2.4 Discussion 10
- 2.3 Risk Factors for Postoperative Patients 12
  - 2.3.1 Background 12
  - 2.3.2 Study Design 14
  - 2.3.3 Results 16
  - 2.3.4 Discussion 18

CHAPTER 3: LEARNING NONLINEAR DISEASE ASSOCIATION NETWORKS 26
- 3.1 Introduction 26
- 3.2 Related Work 29
- 3.3 Proposed Sparse Tree Embedded Graphical Model 30
  - 3.3.1 Formulation 30
  - 3.3.2 Algorithm 32
- 3.4 Numerical Experiments 34
  - 3.4.1 Simulated Data 34
  - 3.4.2 Application on Type II Diabetes Patients 37
- 3.5 Conclusion 39

CHAPTER 4: SURGERY CENTER OPERATING ROOM SCHEDULING 46
- 4.1 Introduction 46
- 4.2 Estimation of Operation Time 49
  - 4.2.1 Data Collection 49
  - 4.2.2 Distribution Fitting 50
- 4.3 Daily Staffing Scheduling 52

4.4 Week-Ahead Stochastic Programming Scheduling 54
4.5 Conclusion 57

CHAPTER 5: CONCLUSION 64

REFERENCES 68

APPENDIX A: COPYRIGHT PERMISSIONS 78

LIST OF TABLES

Table 1 Exclusion criteria for single admissions or patient records 20

Table 2 Significant risk factors across disease groups and factor categories 20

Table 3 Descriptive statistics for risk factors 21

Table 4 Risk ratio values in point estimate 22

Table 5 Significant variables of readmission for colorectal surgery patients 24

Table 6 Comparison of 30-day readmission predictive models 24

Table 7 Nomenclature of daily staffing scheduling model 60

Table 8 Nomenclature of week-ahead stochastic programming scheduling model 61

Table 9 Computational results of goal programming for week 1 63

Table 10 Computational results of goal programming for week 2 63

Table 11 Computational results of goal programming for week 3 63

LIST OF FIGURES

Figure 1 Dissertation outline....3

Figure 2 CTREES produced using identified risk factors on subsample population 25

Figure 3 CTREES produced using identified risk factors on whole population....25

Figure 4 A general framework of regression-based methods 41

Figure 5 Our proposed algorithm for solving (3)....41

Figure 6 Performance comparison on big-n data 42

Figure 7 Performance comparison on big-p data 43

Figure 8 Performance comparison using more general non-linear underlying structures 43

Figure 9 Illustration of networks of non-linear associations 44

Figure 10 Clinical association networks of top 25 diagnosis codes 45

Figure 11 Comparison of empirical and theoretical densities 59

Figure 12 Lognormal distribution information of frequent procedures....59

Figure 13 Descriptive statistics of infrequent procedures....59

Figure 14 Graphical user interface of day-ahead scheduling tool 60

Figure 15 Goal programming for proposed model 62

Figure 16 Scenarios of case samples from week 2 62

ABSTRACT

The efforts to improve health care delivery usually involve studies and analysis of patient populations and healthcare systems. In this dissertation, I present the research conducted in the following areas: identifying patient groups, improving treatments for specific conditions by using statistical as well as data mining techniques, and developing new operation research models to increase system efficiency from the health institutes' perspective. The results provide better understanding of high risk patient groups, more accuracy in detecting disease' correlations and practical scheduling tools that consider uncertain operation durations and real-life constraints.

CHAPTER 1: INTRODUCTION

The goal of improving the health care delivery system is to improve patient outcomes and lower costs. My research focuses on three aspects of healthcare delivery systems: patient populations, detecting clinical associations, and increasing system efficiency. Figure 1 shows the structural outline of the dissertation. Three specific research studies were conducted including hospital readmissions, nonlinear associations of conditions, and scheduling operating rooms (ORs). Various approaches such as regression tests, nonlinear association tree, and optimization models are proposed or applied in these studies.

First, a rate of 30-day hospital readmissions has been established as a hospital's performance measure in promoting quality and patient-centeredness. To have a great impact on improving outcomes with effective cost management, we put effort on complex and vulnerable patient populations, such as individuals with serious chronic conditions, and surgery patients who have high risks of developing surgical complications. To identify the high risk factors and patient population for readmission patient groups, we applied several supervised learning data mining techniques which treat indictor readmission as the response variable. We implemented a multivariate logistic regression model, a proportional hazard model with recurrent events (multiple readmission records), and a conditional tree model.

Second, discovering clinical associations between disease conditions could lead to a better understanding of the readmission risk and guidance for intervention allocation. We developed a novel tree-embedded sparse regression learning graphical model (STGM), which

uncovers both linear and nonlinear relationships from a large number of variables. We further proposed an efficient regression-based algorithm for learning the STGM from data and conducted simulation studies that demonstrated the superiority of the STGM over other network learning methods. We applied our STGM to learn the clinical association networks for readmission analysis in the context of Type-II diabetes that is known for high readmissions rates. The finding shows that certain complications might be the risk factors which increase the complexity of patient conditions.

Finally, optimizing the scheduling of operating rooms (OR) is a challenging problem due to the uncertainty of operation durations as well as material and human resource' constraints. Taking these challenges into account, ambulatory surgery centers (ASCs) must also flexibly adapt to a wide variety of external realities, such as balancing service to providers against efficiency. In this study, we first analyzed the historical data from an ASC, and constructed the lognormal distributions for surgery durations of specialties. Then we proposed a day-ahead nurse staffing model for the surgery center considering the fixed surgery schedules and the affinities between nurses and surgeons. Finally, we formulated a multi-objective stochastic programming model for weekly operating room scheduling. We included the obtained lognormal distributions, and considered the affinities between team members as well as the efficiencies of a nurse assistant during various surgeries.

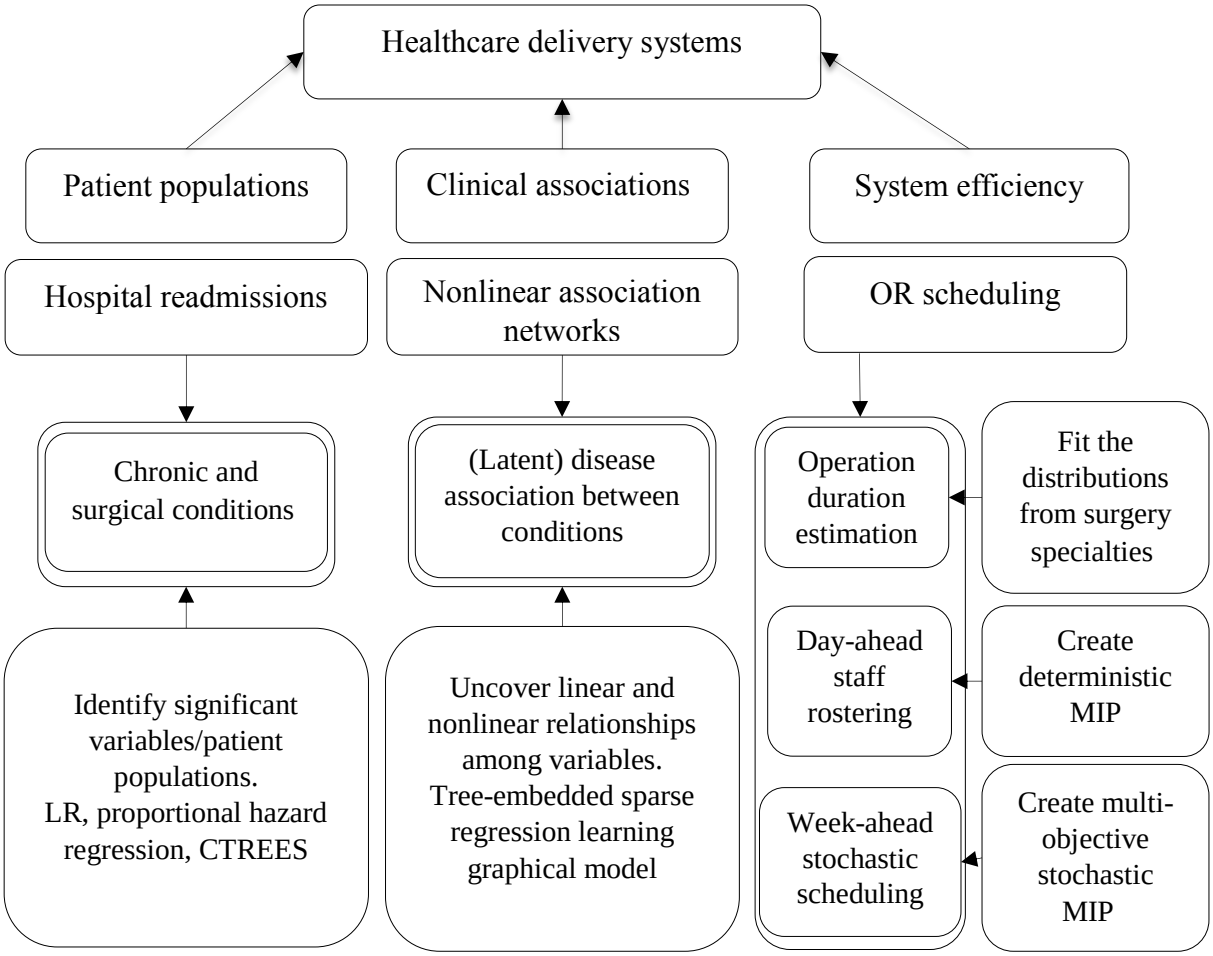

Figure 1 Dissertation outline

CHAPTER 2: RISK FACTORS OF HOSPITAL READMISSIONS[1]

2.1 Introduction

To promote healthcare quality and safety, all-payer 30-day readmission rate has become one of the two performance measures that the National Quality Strategy (NQS) has endorsed [1]. The all-cause readmission rate was about 18 percent by the start of 2013, and about 29 percent of post-hospital home health stays result in readmission [2]. Almost 20% of the 12 million Medicare payers were readmitted after being discharged within 30 days and 34% within 90 days. Only 10% are planned readmissions [3]. The cost of Medicare for readmissions is estimated at $26 billion annually, and more than half of the cost (estimated at $17 billion) is potentially preventable [4].

Centers for Medicare and Medicaid Services (CMS) decreased reimbursements for excessive readmissions recently. For example in 2014, CMS applied algorithms to account for unplanned readmissions for chronic conditions: acute myocardial infarction (AMI), heart failure (HF) and pneumonia. In addition to these three conditions, patients admitted for acute exacerbation of chronic obstructive pulmonary disease (COPD) and patient with admissions for surgery procedures (elective total hip arthroplasty (THA) and total knee arthroplasty (TKA)) are included in the category from 2015[5].

Retrospective analyses have been conducted to identify the risk factors for readmissions. Since the underling medical conditions vary, most studies have been undertaken under different

[1] Portions of this Chapter were previously published in [6] and [21]. Permissions are included in Appendix A.

medical conditions, which produce more interpretive results and further opportunities for interventions on the cohort patient groups. Our study analyzed administrative information of complex and vulnerable patients, specifically patients with serious chronic conditions (in section 2.2) and postoperative patients (in section 2.3).

2.2 Risk Factors for Patients with Chronic Conditions

2.2.1 Background

Patients with chronic conditions contribute the most to readmissions. The underlying causes include poor discharge transitions and instructions, lack of family support, patient complications, and medical error [7]. A number of studies have focused on the common chronic conditions, such as, CHF, COPD, pneumonia, and AMI [8-11], and other disease groups like patients with Type II diabetes or cancer [12, 13]. The readmission prediction models pose difficulty in reaching generalizable results due to different patient study population, sample size, and limited data resources. For example, clinical data is normally not included which could cause some risk factors to be undetectable [14, 15]. However, studying administrative data of high volume hospitals could still help advance the knowledge of causes and factors related to readmissions. The statistical technologies applied in identifying risk variables are very broad in readmission research. The most commonly used technique is logistic regression (LR) [16-24], which is used for predicting binary outcomes of the dependent variables. The 30-day readmission (readmitted vs. not readmitted) is the binary outcome and all risk factors are the predictors in LR model. The underlying assumption of LR is that the transformation of linear combination of predictors is linearly related to the response variable. Another regression model - proportional hazard model (Cox model) has also been implemented to estimate the risk over time. The covariates of Cox model are multiplicatively proportionally related to hazard, and the baseline

hazard function is not necessarily specified. Cox models have been applied to identify significant factors related to readmissions and high risk patient groups [25, 26]. Moreover, other statistical results have been found for readmission studies by implementing univariate analysis and hypothesis testing [27, 28].

2.2.2 Study Design

Our study aims to identify preventable readmissions based on multi-hospital administrative data, and to estimate significant risk factors related to readmission through a multivariate LR model and an extension of Cox model. The results are compared across patient factors, hospital factors, and disease groups.

The original dataset is from a network hospital system consisting of 9 hospitals in central Florida. The dataset includes 7 year retrospective data that includes more than 1 million patient discharge records of about 600 thousands patients from 2005 to 2012. We processed the data in the following steps.

First, the records considered as routine, planned and unavoidable are excluded based on the CMS report. The planned readmissions are those with a pre-specified procedure [29]. Our studies only consider unplanned 30-day admissions. The exclusion criteria are shown in Table 1. After these eliminations, the data contains only preventable readmissions with 470 thousand patient records and 760 thousand hospitalizations.

Second, five common chronic conditions are selected as our study cohorts. The selection is based on primary diagnosis code (ICD-9-CM) [30] of index admissions. The conditions include 1) congestive heart failure (CHF) according primary diagnosis codes 428.*,402.01, 402.91, 404.01, 404.03, 404.11, 404.13, 404.91, and 404.93, 2) chronic obstructive pulmonary disease (COPD) according primary diagnosis codes 491.0, 491.1, 491.2, 491.20, 491.21, 490,

492, and 496, 3) acute myocardial infarction (AMI) according primary diagnosis codes 410.*, 4) Type II diabetes according primary diagnosis codes 250.*2, 5)pneumonia, according primary diagnosis and primary diagnosis related symptom codes 480, 481, 482, 483, 485, 486, 510, 511.0, 511.1, 511.9, 780.6, 780.6, 786.00, 786.05, 786.06, 786.07, 786.2, 786.3, 786.4, 786.5, 786.51, 786.52, 786.7 and a secondary diagnosis codes of pneumonia, emphysema, or pleurisy.

Finally, fifteen potential risk factors are classified into three categories: patient factors, condition severity factors, and hospital factors. The patient factors are patient age (range groups 18 to 45, 45 to 55, 55 to 65, 65 to 75, 75 to 85, 85 and up), gender (Female, Male), marital status (divorced/separated, married, single, widowed), race/ethnicity (black, Hispanic, white, other), and language (English, other). Condition severity factors are severity of illness (1-minor, 2-moderate, 3-major, 4-extreme) as defined by All Patient Refined Diagnostic Related Groups *(*APR DRG)[31], behavioral health comorbidities (Yes if present as a secondary diagnosis, No otherwise), Charlson comorbidity index (0, 1, 2, 3, 4, 5 and up) calculated based on comorbid conditions and their severities [32], and length of stay LOS (in days). The hospital factors are hospitalist (Yes if patient has hospitalist, No otherwise), payer class (commercial, Medicaid, Medicare, other including patients with no insurance), discharge disposition (non-acute facility, routine/home, specialty hospital, other), admission type (emergency, routine, urgent, other), number of previous readmissions, and year (over seven years). The detailed categories of variables and their descriptive statistics are shown in Table 3.

Two statistical models (LR model and a proportional hazard model extension) were built to identify significant variables and the relative risks of patients with different combinations of risk factors. The following is a brief review of the two methods and how they are applied to the readmission problems in this dissertation.

An LR model on 30-day readmission is applied with 15 predictive variables and a binary outcome of readmission (1 if readmitted, 0 otherwise). A patient record is represented by some linear combination of the predictors. The odds ratio (probability of being readmitted over probability of not being readmitted) is equivalent to the exponential function of linear regression expression. The results of relative risks among different class levels are interpreted as log odds. The regression coefficient estimation uses maximum likelihood estimates. Goodness-of-fit for model evaluation is tested by Hosmer-Lemeshow statistic and a 10 fold cross-validation. The results of statistical significant variables are produced by a Wald test with p-value of 0.05.

A proportional hazard model with recurrent events is also applied. The basic Cox model estimates the coefficients for each predictive variable. It uses two responsive variables to capture the risk of event over time. The two responsive variables are event indicator (Y/N) and lapse time (between initial time record and the time when event occurs). The results of LR and Cox regression LR are estimated in different ways that LR estimates the odds ratio while Cox aims to estimate the hazard ratio. The final model from a Cox regression would yield an equation for the hazard as a function of several predictive variables. Since there are patients having multiple records that are readmitted to the hospital multiple times (note that patients returned to the hospital for a totally different reason would not count as readmissions), the patient records are not independent if they are from the same patient, which is called random effects. Also, the readmission events are not independent since patients with multiple previous admissions are more likely to be readmitted. We assume that there are baseline risks for individuals and readmission events. An extended proportional hazard model called conditional frailty model which combines random effects with stratification of events [33], assumes that the effects on the

n^{th} admission event are restricted to the patients who have experienced $n-1^{th}$ admission event. The hazard effect of n^{th} admission for ith patient is:

$$\lambda_{in}(t; Z_{in}) = \lambda_{0k}(t - t_{n-1})e^{\beta' Z_{in}(Q_{in})+\omega_i} \quad (1)$$

where Q_{in} and Z_{in} respectively denote the readmission time (in days) and covariate vector for the i^{th} patient for the n^{th} admission, and β is regression parameter vector. λ_{0n} is the baseline hazard rate for each patient and $(t - t_{n-1})$ represents the gap time (in days) between n^{th} and $n-1^{th}$ admission. ω_i denotes the vector of random effects.

2.2.3 Results

The LR model and hazard regression model were built in SAS and R, respectively. In the 30-day readmission LR predicting model, variable selection is based on a stepwise with settings of entry = 0.1 and stay =0.1. The selection iteratively removes a single insignificant variable from the model and adds a significant variable into the model, until results converge. The results are presented in Table 4 are the odds ratio (OR) for the LR regression model and the hazard ratio (HR) for the hazard regression model, with ratio point and the 0.95 confidence interval (CI).

The results show that significant factors varied across disease groups and were slightly inconsistent in the prediction models. A brief summary of significant factors is shown in Table 2. A number of hospital factors are significant in both models and across disease groups, such as, number of previous readmissions, discharge disposition and year of admission. The more readmitted previously, the more likely it is that the patient would be readmitted again (OR from 1.06 to 1.15). Both OR and HR show decreasing trend of readmission rate over 7 years. Patients who get discharged to another specialty or acute hospital have a higher chance to be readmitted than the ones who are discharged to home or non-acute facilities. Payer class/insurance is identified to be significant for most disease groups, such as CHF, COPD, pneumonia, and Type

II diabetes. The patients with Medicare and Medicaid have higher risks to return to the hospitals than the patients who have commercial insurance. The type of admission is also found significant in most disease groups. More specifically, patients who are admitted as emergency patients have higher risks of readmission. The presence of a hospitalist is not statistically significant in any disease groups.

In the category of case severity factors, length of stay is statistically significant in most disease groups. The longer a patient stays in the hospital, the more likely it is that s/he would be readmitted. Hazard regression model indicates that Charlson comorbidity score relates to the rate of readmission. To be specific, patients with an index of 3 or higher have a higher chance of being readmitted. LR model obtained similar results in patient groups with pneumonia and Type II diabetes. Severity of illness index is significant in patients with CHF, COPD and pneumonia. When the index is high, the odds of being readmitted increases.

Results of patient factors vary across disease groups. For instance, age is found to be significant in Diabetes II patients while hazard model is significant for all patients except patients with AMI. Patients who speak another language have higher risk of readmission in group CHF, COPD and AMI. Patient's marital status also relates to readmission rate. Patients who are divorced are more likely to be readmitted than patients who are married.

2.2.4 Discussion

Our analyses provide information to better understand the risk factors of readmission patients with common conditions. Consistently the study shows that patient discharge disposition, previous admission times, Charlson comorbidity index, length of stay, and insurance type are related to a high readmission rate.

Some interesting results need further investigation. For instance, patients with longer length of stay have a higher risk of readmission. The reason might relate to patient health conditions or infections during hospital stay. With limited information in our data, it is difficult to explain the root cause of this problem. Patients speaking another language are associated with a higher risk rate. We have observed that patients who speak English as the second language are less sick in terms of Charlson Comorbidity Index score and disease severity index. The length of stay for those patients is generally longer (1 day on average) than other English speakers, and it appears that they do not have as many hospitalists. The non-English speaking patients have accounted for the majority of patients with Medicaid, Medicare, and no insurance (less commercial insurance). With these results, however, the conclusion for non-English speaking patients is uncertain. Languages could be a single causable risk factor of hospital readmission. Previous studies suggest that the communication of discharge instructions is important to reduce readmissions [113]. These patients might have difficulty understanding and following the instructions. It might also be a surrogate factor of other hidden variables that are not included in administrative data. In the case of age of Type II diabetes patients, patients 55 years old or younger have a higher risk than the patients who are older than 55. The reason behind this phenomenon might be associated with personal health experience and management. Also, commercial insurance payers have a lower rate of readmission. The results raise a question: is the different readmission rate among payers due to hospital delivery systems or the patient socioeconomic status? Could be both? We do not have more information regarding hospital reimbursement and patient personal information to support any of our suspicions. An in-depth study is required to answer these questions.

Another limitation of our study is that all our data is from hospitals all located in the same metropolitan area in Florida. The results may not represent the whole patient population and hospital characteristics across the country, or even in Florida? The administrative data does not include clinical information and treatments. There is also no information regarding patient discharge and following up information outside the hospital network.

2.3 Risk Factors for Postoperative Patients

2.3.1 Background

As mentioned in section 2.1, in 2015 CMS started to penalize excessive readmissions for two common surgery procedures. For surgical patients, unlike the chronic condition patients, the risk of being readmitted mainly relates to postoperative complications [34, 35]. In the US, more than 50,000 patients had colorectal surgery every year from 1993 to 2007[36]. A systematic review found that the patients who have major bowel surgery account for 16.6% of all patient readmissions for a surgical reason which is a large part of readmissions of Medicare beneficiaries. Mortality associated with colorectal surgery is reported from 1% to 6% in the general population [36, 37]. Besides malignancy, other conditions including inflammatory bowel disease, diverticulitis and diverticular disease, and anal problems often require colorectal surgery [38]. A surgery can normally be either open-or laparoscopic-assisted colectomy, and sometimes with procedure of colostomy. After the surgery, patients could experience different levels of side effects or pain, stress, and dysfunction, which affect patients' quality of life. Although some of the readmissions are unavoidable due to patients' pre-existing complications, a portion of unplanned readmissions are potentially preventable. The causing symptoms for unplanned readmission are bowel obstruction (33.4%), followed by surgical site infection (SSI) (15.7%) and intra-abdominal abscess (12.6%) [39]. Other surgical related reasons are anastomotic leaks,

ostomy-related complications, respiratory complications, etc. and nonsurgical reasons, such as medication complications and side effects of chemotherapy (radiation therapy).

The cost of readmission after colorectal surgery is $9000 each time, a total up to $300 million per year [40]. The readmission rate of patients after colorectal surgery has been increasing over the past 20 years based on a national cancer database [41], while the patient length of hospital stay has been decreasing. In recent years, several intervention strategies have been implemented to surgical patients. A multimodal perioperative intervention care called enhanced recovery pathways (ERP) [42], enhanced recovery after surgery (ERAS) [43], and fast-track surgery [44, 45] have gained widespread acceptance. The program aims to accelerate recovery, shorten patient in-hospital stay, and decrease complications by implementing, for example, optimal pain relief, stress reduction with regional anesthesia, minimal invasive surgery, early nutrition and ambulation. According to several random clinical trial studies, colorectal surgery patients' hospital length of stay (LOS) was significantly reduced in the ERAS group but there was no reduction for major complications and readmission rate [38]. The concern of premature discharge that might increase postoperative morbidity and hospital readmission leads us to pursue standard discharge criteria. However, based on a systematic review of hospital discharge criteria for colorectal surgery [46], there is a huge variety of hospital discharge criteria. That is, a total of 156 studies described 70 different sets of readiness discharge criteria. Thus, development of proper criteria will provide patients a better "ready to discharge" health condition and reduce preventable readmission risks.

In the past, a few studies have tried to identify readmission risk factors/predictors [40, 47-48]. However, most of the risk factors are demographical and social-economic variables which provide little information about the patient health/medical conditions. Even some

contradictions of the risk factors are found, such as patient LOS and hospital volume. These incomplete/inconsistent readmission risk outcomes may result from the limitation of administrative data sets for which some risk factors are not captured by billing codes [47]. Readmission not only relates to postsurgical complications but also involves discharge processes and care coordination [49].

The American College of Surgery National Surgical Quality Improvement Program (ACS-NSQIP) provides risk-adjusted, outcomes-based measures and comparisons for participant hospitals. It generates risk estimates morbidity and mortality probability for individual patients which are created as a function of risk variables. ACS-NSQIP is collected from the patient's medical chart than insurance claims and tracks patients for 30 days after their operation. ACS NSQIP identified 61% more complications including 97% more surgical site infections (SSIs) than administrative data program [50]. Reviewing ACS NSQIP data might provide more insights for readmission problems.

2.3.2 Study Design

We studied 2011 ACS NSQIP data to identify risks of 30 day readmissions in patients that underwent colorectal surgeries. The definition of colorectal surgeries is based on Current Procedural Terminology (CPT) codes (44140 to 44213). A total of 252 variables were included in the original dataset with 30412 records (3228 readmission records, 24370 non-readmission records, 2814 no indicator records). Readmission rate was around 11.7% after eliminating records without readmission indicators, with is congruent with the readmission rate in the literature. To avoid that the information of non-readmission dominates readmission records in data-processing, we limited the size of non-readmission records down to the same size of the readmission records by random subsampling. Consequently, a total of 6456 records were

selected for data analysis (3228 readmission records and 3228 non-readmission records) with a readmission rate of 50%, which provides equal amount of binary outcomes for further analysis.

Dimensionality reduction is a big challenge when preprocessing high dimensional data. Traditional dimensionality reduction methods are unsupervised which only consider input data and exclude any incorporated output data. One of the most popular unsupervised methods is principal component analysis (PCA), which obtains linearly uncorrelated variables from correlated input variables by singular value decomposition of the input matrix. However, ignoring the output/response value would create information loss and bias for discriminant analysis when output information is available. In the case of supervised projection, the relationship between response matrix and observation matrix is assessed. There are some supervised methods, such as PLS-discriminant analysis (PLS-DA) where the outcomes are binary variables. PLS-DA is treated as an optimization problem which maximizes the class separations, based on the covariance between latent variables which are the linear combinations of input variables and response variables [52]. Since our study focuses on analyzing risk factors for readmissions of patients after colon surgeries, the response value is binary (either readmitted or not readmitted). Therefore, PLS-DA could be implemented. Other analytical methods used for classifications could also be applied for identifying risk factors and predicting readmissions, and the most common one is logistic regression (LR) presented in the previous section. Other relatively newer techniques such as support vector machine (SVM) and random forest (RF) [53] have some advantages. For example, they are able to handle high dimensional input variables and are able to find non-linear global solutions. SVM tries to obtain a partition that separates the data well while finding a partition with large margin. RF is a collection of decision trees that are not influenced by each other. The overall prediction of RF is the sum of the predictions from

decision trees. SVM and RF are black-box methods which lack interpretation and could have computational difficulties handling large scale problems. Another interpretable tree structure method called conditional inference tree (CTREE) [54], similar to the traditional decision tree, is also based on recursive partitioning algorithm to determine associations between significant variables. The difference is CTREE uses multiple test procedures to conduct stopping criteria while decision tree applies information measures (such as the Gini index). The latter has a bias of favoring many possible splits or missing values. CTREE selects variables with many possible splits, and applies a statistical test to evaluate the significance of the splits and outcomes. The algorithm recursively selects a covariate, chooses and adjusts the splits until the convergence is reached.

The methods mentioned above were implemented on the subsample ACS NSQIP national surgery records of colorectal patients. The R program provides packages for each method, for example, "glmnet" for Logistic regression, "caret" for PLS-DA, "party" for CTREES, "randomforest" for random forest, and "e1071" for SVM. To identify significant variables and create a split for CTREE, the P value has to be less than or equal to 0.05.

2.3.3 Results

Twenty-seven potential risk factors were identified after data cleaning, and pre-screening by statistical methods (LR) and medical expertise (in Table 5). The factors are categorized into preoperative variables(e.g., patient age and patient previous blood disorder), operative variables (for instance, operation duration), and postoperative variables (e.g., post-surgery complications). The original ACS NSQIP dataset provides the probabilities for mortality and morbidity. The two variables are aggregated from the rest variables.

Our study tested the predictive power of readmission by the probability of mortality probability, probability of morbidity, and the probability of both. The classification accuracy of the three methods shows none or little predicting power. The rates are 0.48, 0.56, and 0.57 (with baseline of 0.5), respectively. Mortality shows no relationship with readmission, while morbidity has a slightly positive correlation with readmission. The combined probability does not increase the predicting accuracy when comparing with morbidity probability, as shown in Table 6. The results of the five analytic methods are also listed in the table. Those well-developed statistical predictive models fail to predict readmissions. 70% classification accuracy is the highest in predicting 30-day readmission with the baseline 50%.

CTREE as a tree structured non-linear classification methods with only a 53% predicting accuracy. However, it is still identified significant sub-groups (branches). CTREE tests the null hypothesis of no relationship between predicting variables and response variable (readmission indictor). If the test fails to reject the null hypothesis, the search ends; If is able to reject the null hypothesis, the predicting variable with the largest correlation value is selected and a split is conducted on that variable. Then the child node is treated as another patient node. The process is recursively repeated until there are no further splits.

The results of the CTREE are shown in Figure 2. The root of the tree (the first split) is organ space surgery side infections (SSIs), and the branches are patients with and without SSIs. The second split on the patients with SSIs is based on hospital length of stay (LOS). And 95% of those patients with LOS less than 10 days were readmitted (443 of 459). Oppositely, the patients with longer LOS (greater than 10 days) have less than 60% (baseline is 50%) readmission rate. Significance test shows another split at LOS (less than 5 days vs greater than 5 days). Note that 13 patients with SSIs stayed less than 5 days had a readmission rate of nearly 100% (220/222).

The patients with SSIs discharged between day 5 and 10 have a readmission rate of 94% (223 of 237).

For the patients who did not develop SSIs, a split leading to the indicator of returning to the OR is created. And the condescending split on the patients who come back to OR is also caused by LOS. The patients who returned to OR and were discharged from hospital within 8 days of index operation had a readmission rate of 93% (252 of 271). Further splits on the patients who neither developed SSI nor returned to OR are caused by the development of superficial infections and American Society of Anesthesiologists (ASA) class.

CTREE was also applied on the whole dataset to evaluate the consistency of the results (shown in Figure 3). The readmission rate of the total population is 11.7%. The tree structure of the whole population is similar to the sub-sample one. The first split is SSIs, and the second split on patients with SSI leads to the duration of hospital stay on 10 days as well. The readmission rate of patients with SSI and stay duration in hospital less than 10 days is 78% (445 of 571). The condescending split is caused by return to OR for patients who did not have SSIs. The difference of the split is the LOS duration. The split point is 12 days instead of 8 days. Patients who returned to OR and got discharged within 12 days have a probability of readmission of 52% (321 of 617). The condescending split of the branch that patients without organ space SSI and did not return to OR is Sepsis. Patients who have Sepsis have a higher readmission rate 30% than other patients (9%).

2.3.4 Discussion

Readmission is an integrated result that could involve various factors. Predicting risk of readmission for individuals is very difficult based on existing data information. As the experiment results shown in our study, typical and advanced analytical methods fall short on

predicting readmission rate. It is also confirmed in the literature that there are few agreements on what are the causal factors. However, our CTREE analysis identifies certain high risk patient subpopulations. It is the first time to apply CTREE on ACS-NSQIP data. The results suggest early discharge could be a main reason for certain patients to get readmitted. Postoperative LOS is significantly related to readmission of patients with organ/space surgical site infection complication and patients who return to operating room within 30 days. It is likely that the infections could cause greater complications, such as anastomotic leak. Also, patients who need to return to OR signify the condition need to be treated. For example, patients developed a superficial SSI or sepsis. A focus only on early discharge following the fast track protocol might be harmful for the patients with serious complications. The discharge plan for those patients should be carefully examined and executed.

The study has the following limitations: it is retrospective and is restricted to colorectal surgery patients. Future work should include a prospective study to investigate additional hospital risk factors. Since administrative data does not tell much detailed information about the patients' experience and the health condition during the hospital stay, investigating the medical records, for example, paired matching groups (like propensity score matching PSM [55]) with opposite outcomes (readmitted vs. not readmitted), will provide a new way to relook at the readmission issue.

Table 1 Exclusion criteria for single admissions or patient records

Admissions The record of the admission (single event) was excluded if it was due to:	**Patients** The entire patient record was excluded if he/she was:
- Continued Care in the same hospital due to same-day internal hospital transfer (This was represented as a readmission in the same day in the database) - Newborn delivery - Trauma - Rehabilitation - Outside transfer and discharge planning is performed - Elopement: leaving without medical advice and/or treatment - Death and subsequent to death (i.e. organ donation)	- Discharged to hospice care - Diagnosed with Cancer: ICD9 code "Malignant Neoplasm" and on-going cancer treatment - Diagnosed with Renal disease and on-going treatment

Table 2 Significant risk factors across disease groups and factor categories

		CHF		**COPD**		**AMI**		**Pneumonia**		**Diabetes II**	
		OR	HR	OR	HR	OR	HR	OR	HR	OR	HR
Patient factors	Age		x		x				x	x	x
	Language	x	x	x		x	x				
	Marital status				x	x		x		x	
	Race		x					x			x
	Gender						x				
Case severity factors	Behavioral health	x									
	Severity of illness	x		x	x			x	x		
	Length of stay	x		x	x			x	x	x	x
	Charlson comorbidity index		x				x	x	x	x	x
Hospital factors	Hospitalist										
	Admission type		x			x				x	
	Payer class		x	x	x			x	x		x
	No. of previous admissions	x	x	x	x	x	x	x	x	x	x
	Year	x	x	x	x	x	x	x	x	x	x
	Discharge disposition	x	x	x	x	x	x	x	x	x	x

Table 3 Descriptive statistics for risk factors

			CHF	COPD	AMI	Pneumonia	Diabetes II
		No. of patients	7287	5946	9688	10897	4879
		No. of admissions	9590	7921	11210	12130	6158
Patient factors	Age	[18, 45)	4.83 (%)	4.61 (%)	6.07 (%)	16.62 (%)	24.90 (%)
		[45, 55)	9.76	14.97	16.88	14.64	22.73
		[55, 65)	13.54	24.07	23.07	14.95	19.31
		[65, 75)	17.02	25.08	19.86	15.34	15.43
		[75, 85)	27.82	21.78	21.08	21.73	12.11
		[85+)	14.93	6.19	7.79	9.32	3.73
		Null	12.10	3.31	5.25	7.40	1.78
	Gender	Female	51.41	56.93	41.28	55.90	49.97
		Male	48.59	43.07	58.72	44.10	50.03
	Marital status	Divorced/Separated	11.29	19.88	10.34	11.83	16.29
		Married	39.74	35.89	51.27	41.28	35.85
		Single	21.30	23.65	22.75	27.13	35.62
		Widowed	27.67	20.59	15.64	19.77	12.24
	Race	Black	15.21	8.98	6.17	11.78	28.28
		Hispanic	8.08	4.94	8.26	8.68	12.85
		White	75.31	84.86	82.40	77.71	56.94
		Other	1.40	1.21	3.17	1.83	1.93
	Language	English	70.22	79.52	78.55	75.19	78.73
		Other	29.78	20.48	21.45	24.81	21.27
Case severity factors	Severity of Illness	1 Minor	9.35	20.26	25.22	10.84	21.60
		2 Moderate	45.29	43.23	40.95	48.41	33.87
		3 Major	35.33	24.25	22.74	31.55	23.22
		4 Extreme	5.52	3.04	9.05	6.10	3.00
		Null	4.52	9.22	2.03	3.10	18.30
	Behavioral health co	No	76.53	65.24	80.09	70.26	74.76
		Yes	23.47	34.76	19.91	29.74	25.24
	Charlson co	0	15.90	0.00	34.87	28.12	10.02
		1	24.59	47.54	31.01	37.00	32.97
		2	22.90	26.70	16.76	18.10	18.27
		3	15.45	12.08	8.18	7.64	15.61
		4	9.69	6.77	4.30	4.43	10.56
		5+	11.47	6.91	4.88	4.71	12.59
	Length of stay(days)	Mean (min,max)	4.6(0,19)	3.8(0,56)	4.1(0,78)	5.2(0,15)	3.8(0,90)
Hospital factors	Hospitalist	yes	25.85	29.10	27.27	28.62	32.64
		no	74.15	70.90	72.73	71.38	67.36
	Payer class	Commercial	9.49	10.96	26.52	18.39	19.96
		Medicaid	10.32	14.47	8.26	12.56	21.14
		Medicare	75.89	67.44	55.98	60.00	44.71
		Other	4.30	7.13	9.24	9.05	14.19
	Discharge disposition	Non-acute facility	43.02	29.57	26.43	33.79	32.49
		Routine/home	52.74	67.10	57.22	63.45	64.08
		Specialty hospital	2.89	1.00	14.99	0.88	0.99
		Other	1.35	2.34	1.36	1.88	2.44
	Admission type	Emergency	83.67	82.07	77.25	87.36	69.29
		Routine	4.53	9.22	2.08	3.10	18.32
		Urgent	6.61	3.64	9.22	4.23	5.31
		Other	5.19	5.08	11.45	5.31	7.08
	No of previous admissions	Mean (min,max)	2.8(1,36)	3.3(1,45)	1.9(1,49)	2.4(1,59)	3.1(1,52)
	Year	05	19.26	13.26	14.89	16.07	14.31
		06	16.03	12.11	13.31	14.55	13.41
		07	13.23	12.02	15.58	13.72	13.30
		08	13.69	14.76	16.33	14.06	14.70
		09	12.40	17.04	14.99	15.00	15.54
		10	14.58	17.28	14.59	15.42	15.85
		11-12	10.81	13.53	10.31	11.19	12.89

Table 4 Risk ratio values in point estimate (0.95 confidence interval)

			CHF		COPD		AMI	
			Odds ratio	Hazard ratio	Odds ratio	Hazard ratio	Odds ratio	Hazard ratio
Patient factors	Age	[18, 45)		1		1		
		[45, 55)		0.94 (0.77, 1.15)		1.52 (1.18, 1.97)		
		[55, 65)		0.78 (0.64, 0.96)		1.6 (1.25, 2.06)		
		[65, 75)		0.73 (0.59, 0.9)		1.46 (1.12, 1.91)		
		[75, 85)		0.78 (0.63, 0.96)		1.27 (0.97, 1.68)		
		[85+)		0.81 (0.65, 1.01)		1.35 (0.98, 1.85)		
	Marital-status	Divorced				1	1	
		Married				0.84 (0.75, 0.95)	1.13 (0.95, 1.36)	
		Single				0.93 (0.82, 1.05)	0.92 (0.75, 1.12)	
		Widowed				0.98 (0.86, 1.13)	1.12 (0.91, 1.39)	
	Race	Black		1				
		Hispanic		0.86 (0.73, 1.02)				
		White		0.81 (0.73, 0.91)				
		Other		0.57 (0.38, 0.85)				
	Language	English	1	1	1		1	1
		Other	1.17 (0.99, 1.38)	1.13 (1, 1.27)	1.27 (1.01, 1.6)		1.19 (1.02, 1.4)	1.13 (0.95, 1.34)
Case severity factors	Disease severity	1	1		1	1		
		2	1.23 (0.99, 1.52)		1.17 (0.97, 1.41)	0.99 (0.88, 1.1)		
		3	1.32 (1.06, 1.66)		1.39 (1.13, 1.72)	1 (0.88, 1.14)		
		4	1.33 (0.97, 1.85)		1.62 (1.09, 2.41)	0.94 (0.71, 1.24)		
	Charlson	0		1				1
		1		1.14 (0.99, 1.3)				1.03 (0.9, 1.19)
		2		1.22 (1.06, 1.39)				1.01 (0.85, 1.19)
		3		1.3 (1.12, 1.51)				1.13 (0.9, 1.42)
		4		1.34 (1.14, 1.59)				1.35 (1.03, 1.78)
		5+		1.26 (1.06, 1.49)				1.03 (0.77, 1.39)
	Length of stay (days)		1.02 (1, 1.03)		1.04 (1.02, 1.06)	1.03 (1.02, 1.04)		
Hospital factors	Payer	Commercial		1	1	1		
		Medicaid		1.36 (1.14, 1.62)	1.94 (1.45, 2.6)	1.56 (1.3, 1.87)		
		Medicare		1.23 (1.04, 1.46)	1.44 (1.11, 1.88)	1.38 (1.16, 1.64)		
		Other		0.87 (0.68, 1.11)	1.55 (1.09, 2.22)	1.48 (1.19, 1.84)		
	Num of admissions		1.15 (1.12, 1.17)	1.08 (1.07, 1.1)	1.15 (1.13, 1.17)	1.09 (1.08, 1.1)	1.12 (1.09, 1.15)	1.14 (1.09, 1.18)
	Discharge dispos.	Non-acute	1	1	1	1	1	1
		Routine	0.83 (0.73, 0.93)	1.05 (0.96, 1.15)	0.9 (0.77, 1.05)	1.04 (0.94, 1.16)	0.6 (0.52, 0.69)	0.74 (0.65, 0.85)
		Specialty	2.43 (1.85, 3.2)	1.74 (1.4, 2.17)	2.13 (1.27, 3.58)	1.45 (0.98, 2.15)	6.74 (5.82, 7.81)	41.1 (33.99,
		Other	1.59 (1.04, 2.44)	1.27 (0.93, 1.72)	1.78 (1.21, 2.62)	1.58 (1.21, 2.06)	1.1 (0.71, 1.72)	1.36 (0.86, 2.16)
	Admission type	Emergency		1			1	
		Other		0.8 (0.65, 0.99)			1.1 (0.78, 1.55)	
		Routine		0.83 (0.7, 0.98)			0.73 (0.58, 0.9)	
		Urgent		0.87 (0.73, 1.04)			0.84 (0.69, 1.01)	
	Year	1	1	1	1	1	1	1
		2	0.88 (0.73, 1.06)	0.86 (0.76, 0.97)	0.96 (0.75, 1.23)	0.91 (0.78, 1.06)	0.85 (0.7, 1.04)	0.86 (0.7, 1.06)
		3	0.77 (0.61, 0.97)	0.83 (0.71, 0.97)	0.88 (0.65, 1.2)	0.84 (0.72, 0.98)	1 (0.81, 1.24)	0.9 (0.71, 1.13)
		4	0.84 (0.66, 1.05)	0.84 (0.71, 0.98)	0.85 (0.63, 1.15)	0.79 (0.68, 0.91)	0.85 (0.69, 1.06)	0.8 (0.64, 1.01)
		5	0.72 (0.57, 0.92)	0.7 (0.6, 0.83)	0.78 (0.58, 1.04)	0.69 (0.6, 0.81)	0.91 (0.73, 1.14)	0.76 (0.6, 0.96)
		6	0.76 (0.6, 0.96)	0.74 (0.63, 0.87)	0.72 (0.53, 0.97)	0.62 (0.53, 0.72)	0.74 (0.59, 0.93)	0.66 (0.51, 0.84)
		7-8	0.57 (0.44, 0.74)	0.43 (0.35, 0.52)	0.54 (0.39, 0.75)	0.3 (0.25, 0.37)	0.73 (0.57, 0.94)	0.56 (0.43, 0.75)

Table 4 (Continued)

			Pneumonia		Type II diabetes	
			Odds ratio	Hazard ratio	Odds ratio	Hazard ratio
Patient factors	Age	[18, 45)		1	1	1
		[45, 55)		1.07 (0.89, 1.27)	1.8 (0.55, 5.87)	1.01 (0.84, 1.21)
		[55, 65)		1.03 (0.86, 1.23)	1.03 (0.31, 3.4)	0.68 (0.55, 0.84)
		[65, 75)		0.84 (0.68, 1.03)	1.52 (0.46, 5.05)	0.67 (0.51, 0.88)
		[75, 85)		0.79 (0.65, 0.97)	1.8 (0.54, 6)	0.73 (0.55, 0.97)
		[85+)		0.83 (0.66, 1.05)	2.11 (0.61, 7.37)	0.65 (0.43, 0.98)
	Marital-status	Divorced	1		1	
		Married	0.77 (0.64, 0.92)		0.82 (0.65, 1.03)	
		Single	0.85 (0.7, 1.03)		0.91 (0.72, 1.14)	
		Widowed	0.72 (0.59, 0.89)		0.62 (0.44, 0.87)	
	Race	Black	1			1
		Hispanic	0.79 (0.6, 1.04)			0.8 (0.64, 1.01)
		White	1.03 (0.85, 1.24)			0.61 (0.34, 1.08)
		Other	0.85 (0.51, 1.39)			0.95 (0.81, 1.1)
	Language	English				
		Other				
Case severity factors	Disease severity	1	1	1		
		2	1.09 (0.86, 1.39)	1.2 (0.99, 1.45)		
		3	1.32 (1.03, 1.7)	1.36 (1.11, 1.65)		
		4	1.55 (1.12, 2.16)	1.35 (1.04, 1.77)		
	Charlson	0	1	1	1	1
		1	1.16 (0.98, 1.36)	1.26 (1.1, 1.44)	0.9 (0.62, 1.3)	0.95 (0.73, 1.24)
		2	1.27 (1.05, 1.53)	1.37 (1.18, 1.6)	1.73 (1.19, 2.5)	1.58 (1.2, 2.07)
		3	1.4 (1.11, 1.77)	1.47 (1.22, 1.78)	2.01 (1.38, 2.91)	1.74 (1.32, 2.29)
		4	1.57 (1.2, 2.06)	1.5 (1.2, 1.89)	1.9 (1.28, 2.83)	1.96 (1.46, 2.63)
		5+	1.55 (1.19, 2.02)	1.56 (1.25, 1.94)	1.87 (1.25, 2.78)	1.67 (1.23, 2.26)
	Length of stay (days)		1.02 (1, 1.03)	1.01 (1.01, 1.02)	1.03 (1.01, 1.04)	1.03 (1.02, 1.04)
Hospital factors	Payer	Commercial	1	1		1
		Medicaid	1.6 (1.26, 2.02)	1.73 (1.44, 2.08)		1.51 (1.23, 1.85)
		Medicare	1.47 (1.21, 1.78)	1.79 (1.5, 2.14)		1.31 (1.06, 1.63)
		Other	1.02 (0.76, 1.37)	1.02 (0.81, 1.28)		1.07 (0.85, 1.35)
	Num of admissions		1.09 (1.07, 1.11)	1.06 (1.05, 1.08)	1.11 (1.09, 1.12)	
	Discharge dispos.	Non-acute	1	1	1	1
		Routine	0.72 (0.62, 0.82)	0.83 (0.74, 0.93)	0.88 (0.72, 1.07)	1.52 (1.09, 2.1)
		Specialty	3.26 (2.14, 4.97)	2.9 (1.98, 4.26)	3.95 (2.21, 7.04)	0.92 (0.8, 1.07)
		Other	1.62 (1.12, 2.35)	1.55 (1.13, 2.11)	2.15 (1.41, 3.29)	3.35 (1.92, 5.85)
	Admission type	Emergency			1	
		Other			0.8 (0.62, 1.04)	
		Routine			0.9 (0.63, 1.27)	
		Urgent			0.73 (0.52, 1.03)	
	Year	1	1	1	1	1
		2	0.96 (0.78, 1.17)	1.01 (0.87, 1.18)	0.73 (0.55, 0.98)	0.68 (0.55, 0.84)
		3	0.89 (0.72, 1.1)	0.89 (0.76, 1.04)	0.65 (0.48, 0.87)	0.83 (0.67, 1.02)
		4	0.91 (0.74, 1.12)	0.95 (0.81, 1.11)	0.63 (0.47, 0.84)	0.71 (0.57, 0.87)
		5	0.76 (0.61, 0.93)	0.77 (0.65, 0.91)	0.59 (0.44, 0.79)	0.59 (0.48, 0.73)
		6	0.76 (0.62, 0.94)	0.74 (0.62, 0.87)	0.51 (0.38, 0.69)	0.52 (0.41, 0.65)
		7-8	0.7 (0.55, 0.88)	0.56 (0.45, 0.68)	0.48 (0.35, 0.66)	0.39 (0.3, 0.51)

Table 5 Significant variables of readmission for colorectal surgery patients

Preoperative variables	Age American Society of Anesthesiologists class ASACLAS History of chronic obstructive pulmonary disease HXCOPD Steroid use STEROID Diabetes DIABETES Bleeding disorder BLEEDDIS Ventilator dependent >48 hours VENTILAT Quarter of admission AdmQtr
Operative variables	Total operation time OPTIME Work relative value units WORKRVU
Postoperative variables	Return to operating room RETURNOR Postoperative length of hospital stay DOptoDis Days from Surgical Admission to Operation HtoODay Superficial infection SUPINFEC Deep Incisional surgical site infection WNDINFD Organ/space surgical site infection ORGSPCSSI Occurrences Bleeding OTHBLEED Occurrences pulmonary embolism PULEMBOL Urinary tract infection URNINFEC Occurrences Ventilator > 48 hours FAILWEAN Occurrences of DVT/Thrombophlebitis OTHDVT Occurrences of Myocardial Infarction CDMI CVA/Stroke with neurological deficit CNSCVA Occurrences of Pneumonia OUPNEUMO Progressive renal insufficiency RENAINF Cardiac Arrest Requiring CPR CDARREST Occurrences Sepsis OTHSYSEP

Table 6 Comparison of 30-day readmission predictive models

Method	Covariate	Classification accuracy
Logistic regression	Mortality probability	0.48
Logistic regression	Morbidity probability	0.56
Logistic regression	Mortality probability + Morbidity probability	0.57
Logistic regression	27 significant variables	0.7
PLSDA	27 significant variables	0.69
CTREES	27 significant variables	0.63
Random forest	27 significant variables	0.69
SVM	27 significant variables	0.69

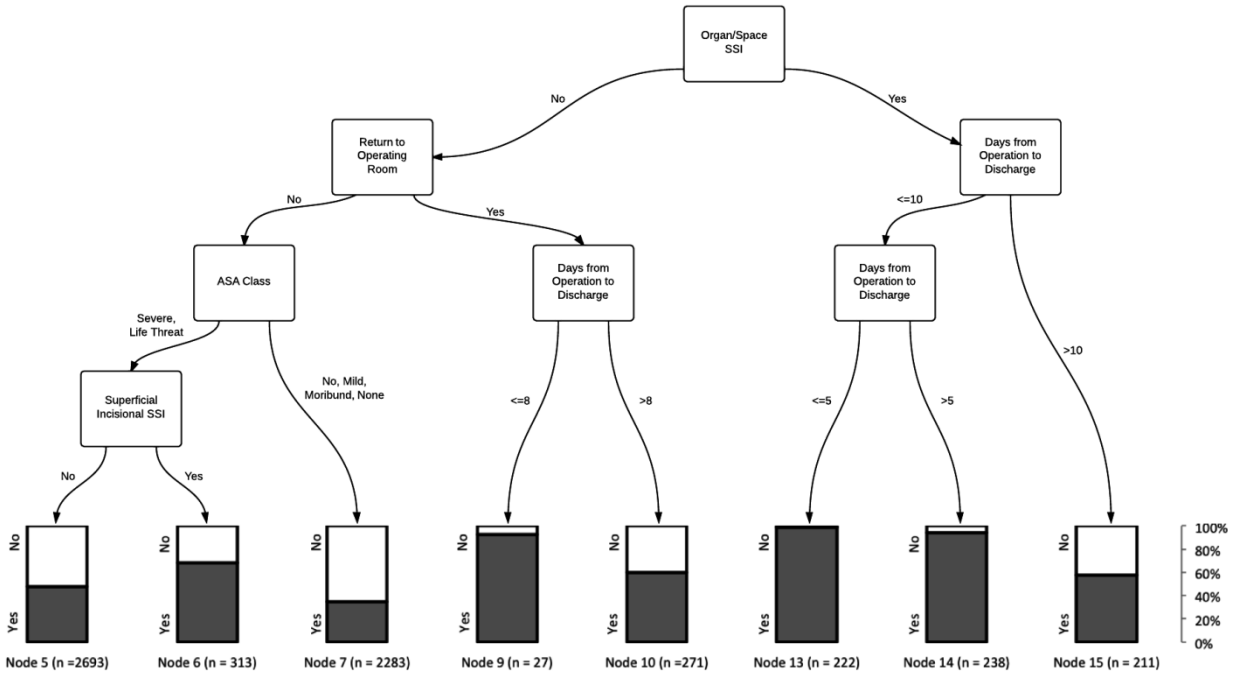

Figure 2 CTREES produced using identified risk factors on subsample population

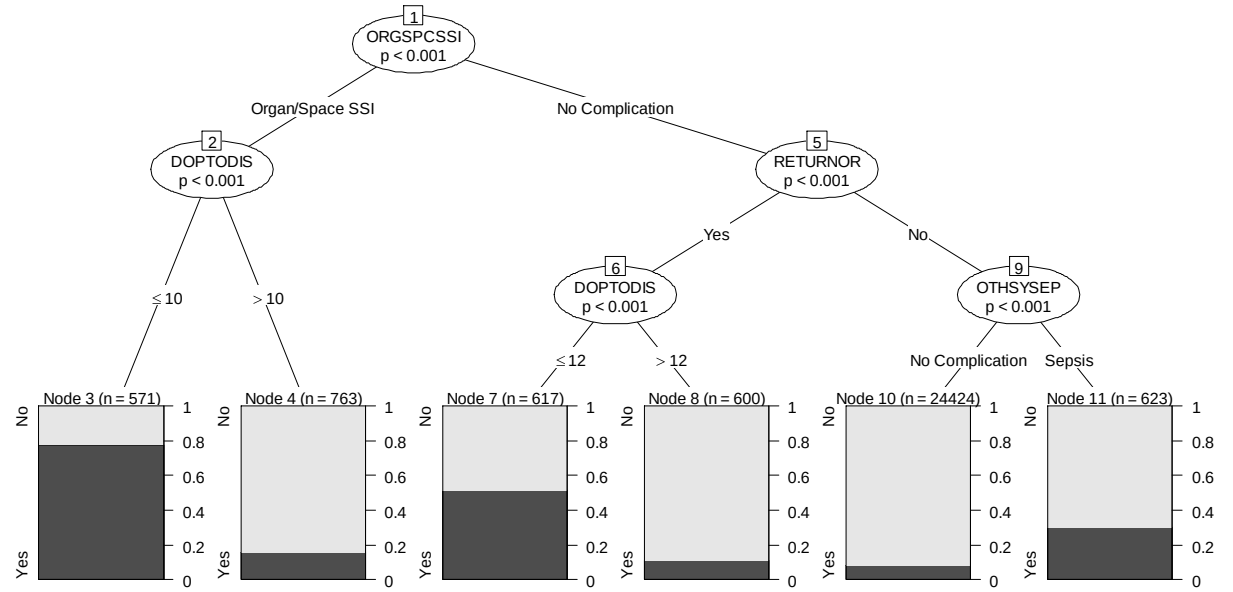

Figure 3 CTREES produced using identified risk factors on whole population

CHAPTER 3: LEARNING NONLINEAR DISEASE ASSOCIATION NETWORKS[2]

3.1 Introduction

Network models have been widely used in many domains to characterize relationships between physical entities. For example, gene association networks have been used to model how different genes interact in a biological process [56]. Brain connectivity networks have been used to model how different brain regions interact to jointly deliver a brain function such as cognition and emotion [57]. Although the networks are not readily measureable in many applications, recent advancement of sensing technologies have risen the possibility of learning these networks from the rich amounts of sensing data, such as gene micro-arrays and brain images for the aforementioned networks, respectively.

Extensive research efforts have been conducted for learning networks from data. Many of them focused on one particular type of network model that is called the Gaussian Graphical Model (GGM). A GGM consists of nodes that are random variables following a multivariate normal distribution and undirected arcs that indicate linear relationships between variables. It has been revealed that learning a GGM is equivalent to estimating the inverse covariance (IC) of the data, because the undirected arcs in a GGM correspond to nonzero entries in the IC matrix of the data [58]. Existing methods for learning a GGM can be broadly categorized as hypothesis-testing-based methods, likelihood-based methods and regression-based methods.

[2]This entire chapter was previously published in [31]. Permission is included in Appendix A.

The hypothesis-testing-based methods employ hypothesis testing techniques to test for each entry of the IC matrix [59-62]. As the number of entries of an IC matrix grows rapidly with respect to the number of nodes, it is difficult to control the overall type-I error since a large number of hypothesis testing will be conducted. As a remedy, the likelihood-based methods were proposed to identify the zero entries in the IC matrix simultaneously. It takes advantage of the assumption that the random variables should follow a multivariate normal distribution in a GGM. Penalized maximum likelihood approaches were proposed in several studies [57, 63-65] that imposed penalties on the entries in the IC matrix, forcing many insignificant entries being zero. Efficient algorithms were proposed by Friedman [63] and Sun [66] to implement the penalized maximum likelihood methods, particularly, for high-dimensional problems. Some other methods were proposed, such as a method based on threshold gradient descent regularization developed by Li and Gui [67], and a method for overcoming the ill-conditioned problem of the sample covariance matrix by Schafer and Strimmer [68]. In addition, there are methods dealing with the situations when variables have a natural ordering [69, 70]. On the other hand, regression-based methods use regression methods for detecting the network structure. For example, Meinshausen and Buhlmann [71] developed a variable-by-variable approach that used lasso regression to identify the neighborhood for each node in the network. Schafer and Strimmer [68] also developed a joint sparse regression model, which simultaneously performs neighborhood selection for all variables. Peng et al. developed a sparse regression technique called SPACE [72], which is particularly useful in identifying hubs in gene association networks. Friedman et al. also investigated the use of lasso and group lasso for fast approximations to exact penalized maximum likelihood estimation of GGM [73]. Their method leads to sparse network estimation that is not only sparse in edges but also in nodes. Recently, Hsieh et al [74, 75] have developed

very efficient algorithms that can remarkably extend the sparse learning of the IC matrix of millions of variables.

Despite the enormous research effort on learning the networks, most of them only focus on linear relationships between variables. For example, a GGM essentially assumes that, the relationship between a variable with the variables that connect with it can be characterized as a linear regression model However, in many applications, both linear and nonlinear relationships will exist between the variables. For example, a particular problem we are studying is the detection of the clinical association networks, which characterize the associations between multiple clinical conditions. Failing to uncover these clinical associations may hinder clinicians from detecting important symptoms, potentially leading to inadequate health care such as inappropriate usage of procedures or insufficient treatments. On the other hand, it is very challenging to identify those clinical associations due to their complicated natures [76].

To tackle the challenge of detecting nonlinear relationships in a network, we developed a novel graphical model, the sparse tree-embedded graphical model (STGM), which is able to uncover both linear and nonlinear clinical associations from a large number of variables. While the term "nonlinear association" can take many possible forms, we focused on a particular type of nonlinear associations that can be characterized by tree models. The basic idea of our STGM is integrating regression-based methods with decision tree learning, since decision tree has been demonstrated to be a powerful tool for learning nonlinear interactions between variables, with no additional cost of increasing the model complexity due to its nonparametric nature. We further propose an efficient regression-based algorithm for learning the STGM from data.

3.2 Related Work

In this section, we will briefly review the related work in the existing methods for learning networks, particularly, in the regression-based methods since our method falls into this category [77, 71, 72]. We use $\mathrm{X} = \{\mathrm{X}_1, \cdots, \mathrm{X}_\mathrm{p}\}$ to denote the p random variables under study. A graphical model of X assigns one node for each X_i and connects two nodes if there is association between them. The structure of a graphical model can be characterized by a $\mathrm{p} \times \mathrm{p}$ adjacency matrix G, with entry $\mathrm{G}_{\mathrm{ij}} = 1$ representing an arc between X_i to X_j and $\mathrm{G}_{\mathrm{ij}} = 0$ otherwise.

The structure learning of the graphical model is equivalent to the identification of the nonzero elements in the adjacency matrix G. Particularly, the regression-based methods decompose the learning problem into p sub-problems, while each sub-problem concerns the identification of the neighbors of a variable. For example, in GGM, the associations between variable X_i with other variables can be modeled as a linear regression model, such as $X_i = \beta_i{}^T X_{/i} + \varepsilon_i$, where $X_{/i}$ denotes all the variables except X_i and ε_i denotes the residual term which is modeled as a normal distribution. The regression-based methods repeatedly use some variable selection models for each X_i and identify the non-zero regression coefficients in β_i [77, 71-72]. The zero regression coefficients in β_i correspond to the variables that are not associated with X_i. A general framework for these algorithms is shown in Figure 4. Here, $\mathrm{f}_\mathrm{i}(\beta_\mathrm{i})$ could be a loss function that encourages many elements in β_i to be zero, i.e., the loss function used in Glasso [63]. The regression-based methods can also be applied to other networks rather than GGM. For example, in some Markov graphical models [77] which model discrete variables, the associations between nodes can be modeled as a logistic regression model if the variables are binary, such as $\Pr(X_i = 1) = g(\beta_i{}^T X_{/i}) + \varepsilon_i$, where g denotes the logit link function and ε_i denotes the residual term which is modeled as a binomial distribution.

The regression-based methods are demonstrated to be computationally efficient and accurate by both theoretical analysis and extensive simulation studies. However, many of them are limited to the applications where the associations between variables are linear. A few studies have attempted to relax these constraints and extend graphical models to capture nonlinear associations [78, 79]. For example, Lafferty [78] proposed two approaches, one made a distributional restriction through the use of copulas as a semiparametric extension of the Gaussian distribution, another one used kernel density estimation and restricted the underlying graphs to be trees or forests. Apparently, these restrictive assumptions limit their applicability in many real-world cases. This is particularly true in many clinical association studies where the nonlinear associations are usually non-smooth and take a rule-based semantics, while the methods proposed restrict the nonlinear associations to be represented as smooth functions.

3.3 Proposed Sparse Tree Embedded Graphical Model

3.3.1 Formulation

As mentioned in section 3.2, the regression-based methods employ a regression model to characterize the associations between variable X_i with its neighbors. The STGM characterizes these associations between variables by integrating the generalized linear regression model with decision tree models, such as:

$$E(X_i) = g\left(\beta_i{}^T X_{/i} + \gamma_i T_i\left(X_{/i}\right)\right) \tag{2}$$

where g is the link function that depends on the type of X_i, $T_i\left(X_{/i}\right)$ is a decision tree model which uses $X_{/i}$ as the input, and γ_i measures the effect of $T_i\left(X_{/i}\right)$ [80]. The rationale beyond this model is to model the linear effects and nonlinear effects separately, which owns better interpretability than black-box nonlinear regression models. Also, a tree-based model is able to capture significant associations between multiple variables in a parsimonious and nonparametric way, while others need

more free parameters to represent these associations. Moreover, this association model will bring in computational advantage as inherited from the computational convenience of the existing tree learning algorithms.

The association model (2) provides the basis for statistically inferring which variables are associated with X_i. This can be done by identifying the variables having nonzero values in β_i and the variables that are selected as inputs in $T_i(X_{/i})$. To discard the variables that are not significantly associated with X_i according to the association model (2), the STGM employs the sparse learning technique [81] that leads to the following optimization formulation:

$$\{\hat{\beta}_i, \hat{\gamma}_i, \hat{T}_i\} = \min_{\beta_i, T_i} - \sum_{j=1}^{n} l\left(x_{ij}, g\left(\beta_i{}^T x_{/i,j} + \gamma_i T_i(x_{/i,j})\right)\right) + \lambda_i \sum_{l \in X_{/i}} |\beta_{il}| + \alpha_i |T_i(x_{/i,j})| \quad (3)$$

Here, let x_{ij} denote the j^{th} sample for X_i and $x_{/i,j}$ denotes the j^{th} sample for all the variables except X_i. n is the sample size. The first term in the objective function, $\sum_{j=1}^{n} l\left(x_{ij}, g\left(\beta_i{}^T x_{/i,j} + \gamma_i T_i(x_{/i,j})\right)\right)$, is the likelihood of the generalized regression model to measure the model fit. In the second term, $\sum_{l \in X_{/i}} |\beta_{il}|$ is the sum of the absolute values of the elements in β_i and thus is the so-called L1-norm penalty. Moreover, the term $|T_i(x_{/i,j})|$ denotes the number of terminal nodes in the tree $T_i(x_{/i,j})$ that is used to regularize the estimation of the tree [82]. The regularization parameter λ_i controls the number of non-zero elements in the solution to β_i, $\hat{\beta}_i$; the larger the λ_i, the fewer nonzero elements in $\hat{\beta}_i$. Because fewer nonzero elements in $\hat{\beta}_i$ correspond to fewer arcs in the learned network, a larger λ_i results in a sparser structure. A similar role is also played by α_i. A benchmark practice in literature is to select λ_i and α_i by cross validation. By solving (3) for each X_i, the associated variables for each X_i can be identified as the non-zero values in β_i and the variables

that are selected as inputs in $T_i(X_{/i})$, and thereby, the network structure is identified. We propose our algorithm for solving (3) in section 3.3.2.

3.3.2 Algorithm

To solve (3), we propose an iterative fitting algorithm, which builds on the existing optimization algorithms that solve sparse regression models [83, 81] and decision tree learning [80]. The basic idea is motivated from the observation that, with the estimated $\hat{T}_i(x_{/i,j})$, (3) is a standard formulation for sparse generalized linear regression model

$$\{\hat{\beta}_i, \hat{\gamma}_i\} = \min_{\beta_i} -\sum_{j=1}^{n} l\left(x_{ij}, g\left({\beta_i}^T x_{/i,j} + \hat{\gamma}_i \hat{T}_i(x_{/i,j})\right)\right) + \lambda_i \sum_{l \in X_{/i}} |\beta_{il}| \quad (4)$$

It can be solved by many existing algorithms [81, 83]. For example, if the Gaussian link function is used for$g$, thus, we can adopt the LASSO algorithm for solving (4). On the other hand, with the estimatedβ_i, the optimization of (3) can be simplified to the optimization of

$$\hat{T}_i = \min_{T_i} -\sum_{j=1}^{n} l\left(x_{ij}, g\left({\hat{\beta}_i}^T x_{/i,j} + \hat{\gamma}_i T_i(x_{/i,j})\right)\right). \quad (5)$$

We show that the optimization of (5) can be well addressed by adopting the standard framework of decision tree learning [80]. As the learning of a decision tree is repeatedly splitting a node into two child nodes, it relies on a measure of goodness for choosing a candidate split. (5) introduces such a measure of goodness. Specifically, assuming that the tree has grown and now the learning algorithm is probing the splitting scenario on one of the leaf nodes. With a little abuse of notation, denote the samples within this node as$\{x_1, x_2, \dots, x_m\}$, i.e., $\{x_1, x_2, \dots, x_m\}$ could be a subset of the n training samples. Assume that the candidate split will partition the samples into two subsets, $\{x_1, x_2, \dots, x_{m_1}\}$ and $\{x_{m_1+1}, \dots, x_m\}$. Then, we can define the goodness of this split as the reduction of the negative likelihood value, which is

$$-\sum_{j=z_1}^{n_{zm}} l\left(x_{ij}, g\left(\hat{\beta}_i{}^T x_{/i,j} + \hat{\gamma}_i T_i(x_{/i,j})\right)\right) - \left(-\sum_{j=z_1}^{n_{m_1}} l\left(x_{ij}, g\left(\hat{\beta}_i{}^T x_{/i,j} + \hat{\gamma}_i T_i(x_{/i,j})\right)\right) - \sum_{j=z_{m_1+1}}^{n_{zm}} l\left(x_{ij}, g\left(\hat{\beta}_i{}^T x_{/i,j} + \hat{\gamma}_i T_i(x_{/i,j})\right)\right)\right). \quad (6)$$

With such a definition of the goodness, we can adopt the existing decision tree learning algorithms by replacing the goodness measures that were used, such as gini index or entropy index [80], to learn the decision tree $\hat{T}_i$ which optimizes (6). Specifically, in our study, the RPART routine [84] is used for the tree learning that can be implemented in R environment.

As a summary, the overall framework of the iterative algorithm that estimates $\{\hat{\beta}_i, \hat{\gamma}_i, \hat{T}_i\}$ is shown in Figure 5. A naive way for estimating the initial values for $\{\beta_i{}^0, \gamma_i{}^0, T_i{}^0\}$ can be that, first, we estimate an ordinary decision tree on the dataset using existing decision tree learning method, with X_i as the response variable and the other variables as predictors. Denote this tree as $T_i{}^0$. Then, we estimate $\{\beta_i{}^0, \gamma_i{}^0\}$ by employing the sparse regression model with$\hat{T}_i = T_i{}^{t-1}$. The learning algorithm will continue updating the parameters until no further changes are observed on the parameters, e.g., when$T_i{}^t = T_i{}^{t-1}$. Simulation studies performed in the later section demonstrated that this algorithm usually takes only a few iterations to converge.

We would like to conclude this section with some theoretical discussion of the proposed iterative algorithm in Figure 5. Note that our methodology consists of an iterative use of LASSO and standard tree learning. Both LASSO and tree learning have been extensively studied in the literature that revealing conditions to ensure their consistency or rates of convergence to the true underlying model. For example, it has been found that Lasso is consistent on the model selection both in the classical fixed p setting and in the large p setting as the sample size n gets large, when the Irrepresentable Condition is met. Various error bounds have also been derived for decision tree learning that showed promising results on the consistency of the decision tree learning [85-

88]. Results from the Vapnik-Chervonenkis theory suggest that the amount of training data should grow at least linearly with the size of the decision tree [89, 90]. Further research revealed that that the error rate is more closely related to the "effective number of leaves than the number of leaves [86]. Exploration of how these theoretical properties that can be kept in our formulation could be an interesting future research topic. It is reasonable to believe that our method will inherit the nice theoretical properties of both LASSO and tree learning. On the other hand, regarding the computational complexity, it has been known that the computational complexities of LASSO [63] and some decision tree learning methods [91] are both $O(pn)$. Since our iterative algorithm usually terminates in a couple of iterations, the overall computational complexity of our algorithm should be $O(p^2n)$.

3.4 Numerical Experiments

3.4.1 Simulated Data

In this section, we compare our proposed STGM method with existing methods for learning networks. Among the numerous methods that were developed in the literature, Glasso has become the benchmark method for its good performances in various applications. What is more, Glasso can be implemented using the R package that is publically available. Since the performances of both the Glasso and the STGM are related to the parameters that need to be tuned, we use the Receiver Operating Characteristic (ROC) curve to characterize the performance of each algorithm since it can provide a full picture of how each algorithm can perform under different choices of the parameters. A ROC curve shows the sensitivity versus the specificity that can be achieved at various settings for the parameters of the algorithms. Sensitivity measures the proportion of underlying arcs (i.e., the non-zero entries in the adjacency matrix G) that are correctly identified as such. Specificity measures the proportion of negatives

(i.e., the zero entries in the adjacency matrix G) that are correctly identified as such. The algorithm that can achieve the largest area under the curve (AUC value) of the ROC curve is the best one.

The algorithms are compared across various settings, with respect to the number of variables ($p = 50,100,200$), the level of sparseness (e.g., the number of nonzero entries in G/the number of free entries in G = 1/3), and the sample sizes ($n = 500,1000$). For each combination of the levels of these three parameters, we simulate data using three steps. Firstly, we randomly generate the network structure by creating a $p \times p$ adjacency matrix G that has the required level of sparseness. Secondly, we randomly generate the linear effects between X_i with its neighbors in the network, and generate samples from the network with only linear effects. Thirdly, we randomly generate the nonlinear effects between X_i with its neighbors in the network, and add the nonlinear effects to the samples generated from the second step. Specifically, in the second step, we randomly assign linear effects to the neighbors of X_i by randomly generating the corresponding regression parameters using a uniform distribution $Uni\ (\theta)$ with $\theta = [-1, -0.5] \cup [0.5,1]$. Since a network with only linear relationships between variables is a GGM, samples can be generated from a GGM using approaches described in the literature such as in [72]. Then, in the third step, we randomly assign the nonlinear effects to the neighbors of X_i, and generate a tree model for each variable that is nonlinearly related to X_i. To avoid generating too trivial or too complicated tree models, we generate a tree model with depth being two. The cut-off values being used in the tree are generated randomly. Then, using the association model between the variable X_i and its neighbors, nonlinear effects of X_i can be generated which can be added to the samples of X_i generated from the second step.

With the samples collected from the underlying network, the STGM and the Glasso can be applied on these samples, and a ROC curve can be computed for each of the algorithms. To reduce sampling variation, for each simulation scenario, we repeat the simulation for 100 times and a mean ROC curve can be generated for each algorithm. The simulation results are shown in Figure 6. Apparently, the proposed STGM method outperforms the Glasso method in each scenario, i.e., the AUC value of the ROC curve of STGM is larger than the AUC value of the ROC curve of the Glasso. It is apparent that when the dimension p increases, the performance gain of our method in comparison with Glasso also increases. This trends also hold for the setting where p>n, shown in Figure 7, when p=200 and n=100.

Besides conducting simulations on networks where the nonlinear associations can be characterized as tree models, we also evaluated the performance of the STGM method on networks with more general nonlinear interactions. Particularly, we adopted a similar approach developed in Friedman and Popescu [92] to parameterize these nonlinear terms, that involve three-variable interactions term $\prod_{i=1}^{3} \exp\left(-3\left(1-x_j\right)^2\right)$, two-variable interaction term $\exp\left(-2\left(x_i-x_j\right)\right)$, since function $\sin^2(\pi \cdot x_i)$, and linear term. We denote these four types of associations as Type 1-4, respectively. Then we randomly simulate a network with 50 nodes (an example of those simulated networks can be found in Figure 9(a)) using these associations and generate 500 samples for each network. By following the same simulation procedure as described before, the obtained simulation results are shown in Figure 8. The simulation results demonstrate that the proposed STGM is better on detecting those nonlinear associations in general, although it uses a tree based approach for capturing the nonlinear associations. To obtain more details, a "snapshot" of our simulation is also presented in Figure 9(b-c), i.e., we randomly select one trial of the simulations, and present the underlying network as well as the identified networks by both the STGM and Glasso with the

regularization parameter tuned by cross-validation. It indicates that the STGM is particularly advantageous on detecting the 3rd type of associations, while this type of association is more complicated and more difficult for linear models to capture.

3.4.2 Application on Type II Diabetes Patients

We apply our STGM to learn the clinical association networks for readmission analysis in the context of Type-II diabetes which is known for high readmissions rates. Recent studies have revealed the possibility of mining the Electronic Medical Records (EMRs) to discover clinical associations between disease conditions. For instance, a large scale clinical association study has been performed by Hanauer et al.[76] to identify clinical associations from the ICD-9 codes (International Classification of Diseases, Ninth Revision, Clinical Modification). The ICD codes are based on the International Statistical Classification of Diseases and Related Health Problems (commonly abbreviated ICD), which cover many clinical conditions, such as classifications for signs, symptoms, abnormal findings, complaints, social circumstance, and causes of injury or disease. By analyzing these ICD codes routinely collected on a large pool of patients, it is possible to discover clinically relevant associations that may not have been noticed by individual clinicians as it is difficult for them to manage all the details in routine practice.

The data used in this study comes from the administrative data from a certain area health system composed of a network of several hospitals. We identified 4879 patients and 6158 hospitalizations that have the primary diagnosis as Type-II diabetes. To identify the readmission-related clinical associations that may increase readmission risk, we plan to conduct a group comparison study by identifying the clinical association networks of both the readmission group and the control group. Unplanned admissions within 30 days of discharge were considered as the readmission group [93] and the remaining readmissions were considered as the control group. As

most of the ICD-9 codes are only sparsely observed in these records, in this study, we only focus on the clinical associations between the top 50 ICD-9 codes that are most frequently observed in our Type II diabetes cohort including both groups.

The clinical association networks for the readmission group and the control group, learned by Glasso and STGM, are shown in Fig. 10 (a-b), respectively. In the learning of each network, the regulation parameter is chosen by a 10-fold cross-validation. By comparing the two networks, potential insights that may help reduce readmissions can be obtained, which will lead to better understanding of the readmission risk and guidance for intervention allocation. The STGM model might provide even more association information that could not be detected by the Glasso model.

For example, an overall observation is that both groups share many common clinical associations that are documented in existing literature. E.g., it has been found that the development of coronary artery disease is highly related to some diseases such as Diabetes mellitus, hypertension and hyperlipidemia [94], which explains the common association between the ICD-9 codes 272.4 (Hyperlipidemia) with 414.01 (Coronary Atherosclerosis of native coronary artery), and between 272.4 (Hyperlipidemia) with 401.9 (Unspecified essential hypertension). Our study also revealed that 357.2 (Polyneuropathy in diabetes) is associated with 682.7 (Cellulitis and Abscess of foot), 731.8 (Osteopathies), and 707.15 (Ulcer of other part of foot), which is consistent with the existing knowledge that states that these symptoms are common in diabetes patients who are prone to have infection and neuropathic osteoarthropathy [95]. The clinical associations that present in the control group but are missed in the readmission group might either indicate a difference on group characteristics or misdetection of related symptoms in the readmission group. For example, in readmission group, Hypopotassemia

(276.8) has fewer connections with other conditions such as the Hyposmolality and/or Hyponatremia (276.1), and Diabetes with Ketoacidosis (250.12). The clinical associations that only present in the readmission group might indicate risk factors for early readmission. For example, there are significantly more associations between the Tobacco use disorders (305.1) with other ICD-9 codes in the readmission group. This finding suggests that Tobacco use disorder may be a risk factor in the readmission group since it increases the complexity of the clinical conditions by interacting with other conditions [96]. Knowledge of these associations will provide decision support to clinicians to ensure that the important clinical associations are not ignored. In a summary, our model provides a powerful tool for identifying readmission-related associations. Note that our study only reveals associations. It does not imply causation, nor does it prove medical relevance.

3.5 Conclusion

Extensive research efforts have been conducted for learning networks from data; however, many of them were developed for learning networks with linear relationships. In this chapter, we developed a novel graphical model, the sparse tree-embedded graphical model (STGM), which is able to uncover both linear and nonlinear relationships from a large number of variables. The STGM characterizes the relationships between variables by integrating the generalized linear regression model with decision tree models, modeling the linear effects and nonlinear effects separately, which leads to better interpretability and more simplicity than black-box nonlinear regression models. Simulation studies are conducted with respect to various settings, which show that the STGM model outperforms other network learning methods. It is also applied on a real-world application, which demonstrated its efficacy on discovering interesting nonlinear relationships in practice. Future research directions include the investigation of how to extend

STGM to discover more kinds of nonlinear relationships, how to integrate STGM with domain knowledge regarding the relationships between variables, and how to discover directed nonlinear relationships in Bayesian Networks. Also, it will be interesting to extend our network learning method to time-varying networks by using the fussed lasso to characterize the temporal transitions between networks, as suggested [97]. One challenge is how to characterize the dynamics of the nonlinear interactions by developing a temporal tree model. Furthermore, it will be very interesting to investigate the simple screening rules that may facilitate the learning of the network structure. It has been found that a simple method based on univariate screening of the elements of the empirical correlation matrix can achieve comparable or even better performance than many more complex network learning methods [73]. Screening approaches have generated a great promise for sparse learning, e.g., to exclude the irrelevant variables that are guaranteed to have zero coefficients in a regression model, as demonstrated in recent works [98, 99].

Input: sample matrix, X; number of variable, p; regularization

For $i = 1,2, \dots, p$,

optimize $f_i(\boldsymbol{\beta}_i)$ and get $\boldsymbol{\beta}_i$;

End for

Output: $\boldsymbol{\beta}_i$ for $i = 1,2, \dots, p$

Figure 4 A general framework of regression-based methods

Input: samples $\{\boldsymbol{x}_{ij}, j = 1, \dots, n\}$and $\{\boldsymbol{x}_{/i,j}, j = 1, \dots, n\}$; regularization parameters, λ_i; initial values for $\{\boldsymbol{\beta}_i{}^0, \gamma_i{}^0, T_i{}^0\}$; stopping criterion, ϵ.

Initialize:

Let $converge = false$;

Let $t = 1$;

Repeat

$\{\boldsymbol{\beta}_i{}^t, \gamma_i{}^t\}$ is estimated by solving (3) with $\hat{T}_i = T_i{}^{t-1}$ using the existing LASSO algorithm

$T_i{}^t$ is estimated by using the decision tree learning algorithm (4) with (5) as the goodness criteria, where $\{\hat{\boldsymbol{\beta}}_i, \hat{\gamma}_i\} = \{\boldsymbol{\beta}_i{}^t, \gamma_i{}^t\}$.

If

$$T_i{}^t = T_i{}^{t-1}$$

Then

$converge = true$;

Else

$converge = false$;

End if

Let $t = t + 1$;

Until $converge = true$

Output: $\{\boldsymbol{\beta}_i{}^t, \gamma_i{}^t, T_i{}^{t-1}\}$.

Figure 5 Our proposed algorithm for solving (3)

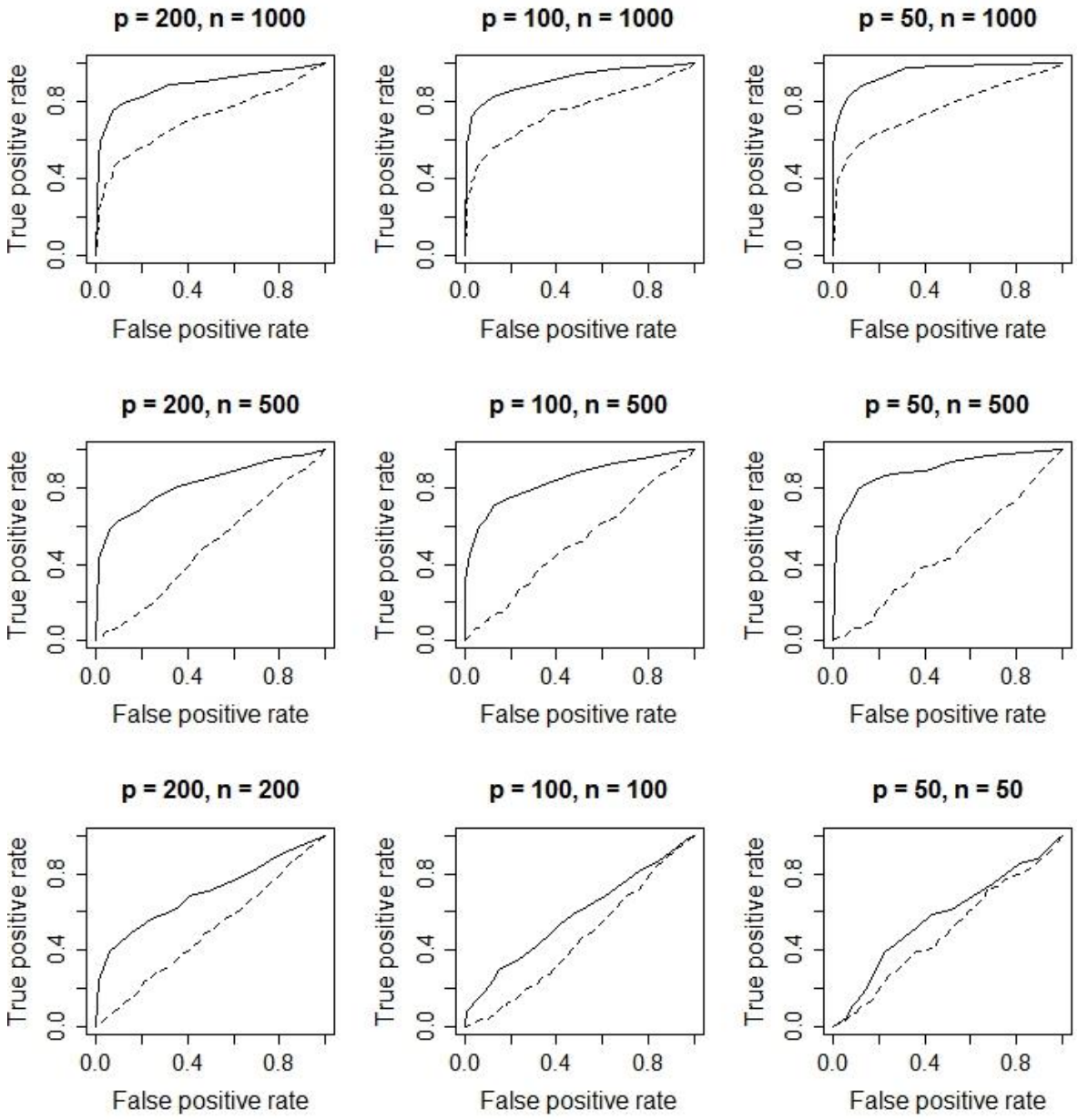

Figure 6 Performance comparison on big-n data. (Mean ROC curves for proposed model (solid line) and the Glasso method (dashed line))

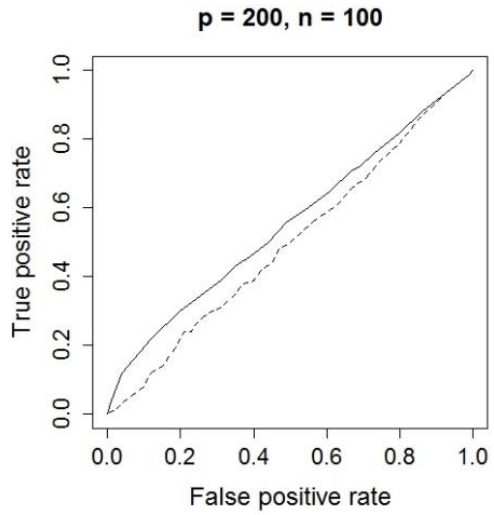

Figure 7 Performance comparison on big-p data. (Mean ROC curves for proposed model (solid line) and the Glasso method (dashed line))

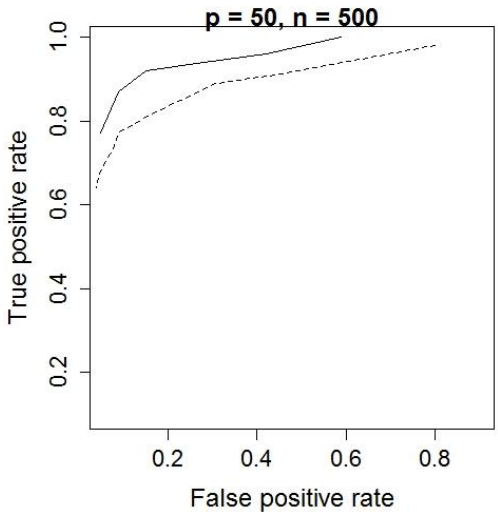

Figure 8 Performance comparison using general non-linear underlying structures. (Mean ROC curves for proposed model (solid line) and the Glasso method (dashed line))

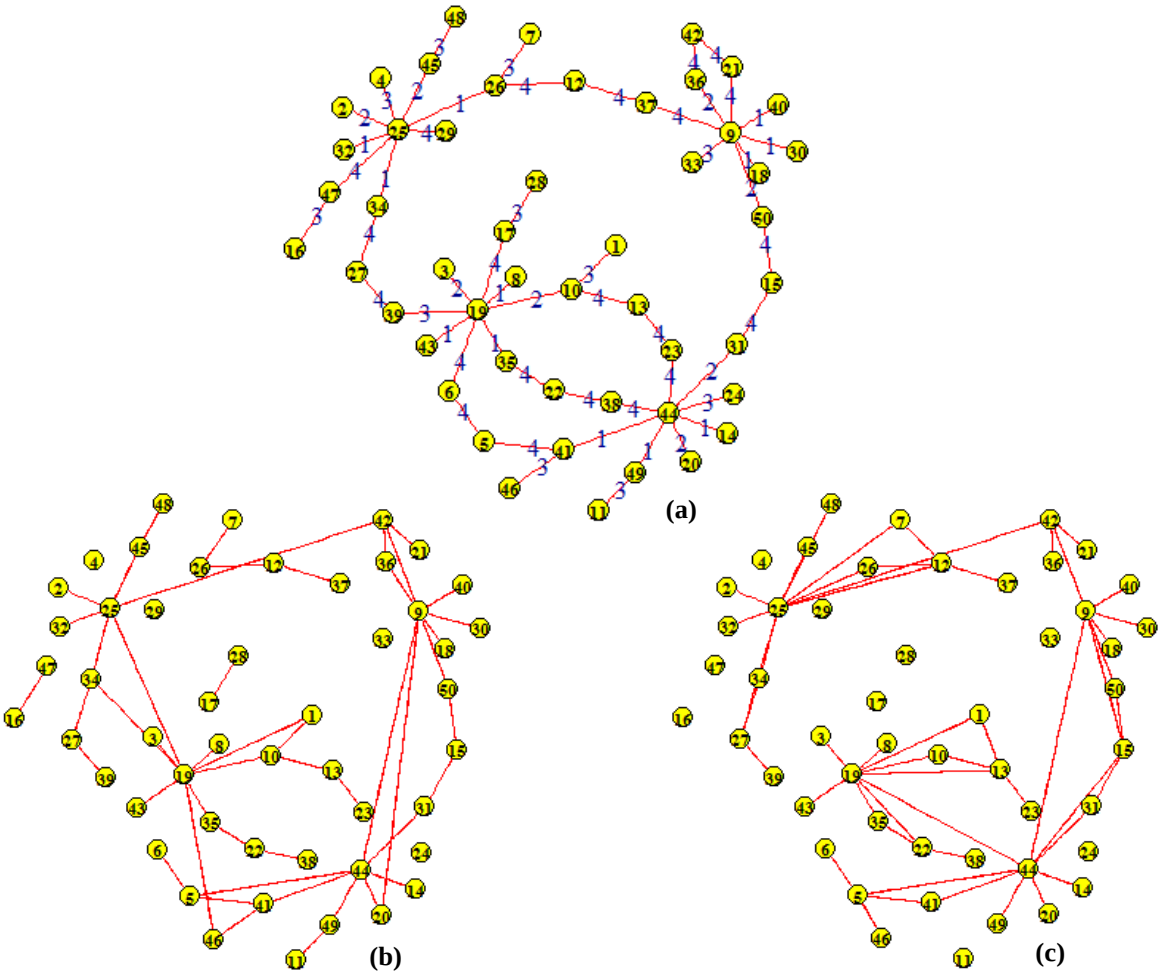

Figure 9 Illustration of networks of non-linear associations. ((a) Real network structure of 50 non-linear associated variables (denoted by vertex) with 4 types of associations (denoted by edge)), (b) Network detected by STEM method, and (c) Network detected by Glasso method)

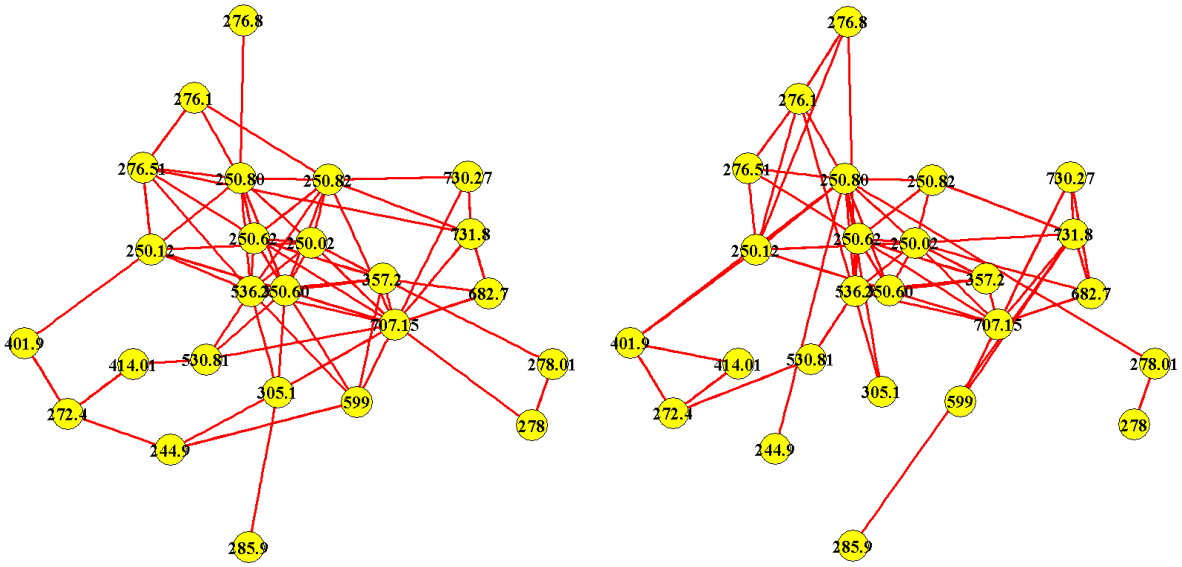

Figure 10 Clinical association networks of top 25 diagnosis codes. ((a) readmission group by STEM method, (b) control group by STEM method)

CHAPTER 4: SURGERY CENTER OPERATING ROOM SCHEDULING

4.1 Introduction

Operating rooms are in general a significant contributor to overall operation cost of health institutes [100]. In recent years, operating room scheduling process is facing more and more challenges and pressures, partly due to higher expectations of patient experience as well as the increasing cost of healthcare delivery. In fact, the difficulty of operating room scheduling resides in the heavy human involvement: surgeons, nurses, patients, and their interactions. All of those factors make this process very different from other scheduling problems such as those in supply chain or electrical systems where physical laws apply.

For the sake of history, we have to mention [101,102] which were published in 1968 and, to our best knowledge, were the first two literatures trying to solve operation room scheduling problems with computer programs. In [101], a simulation program was used to evaluate various schedules for operating rooms by reporting idle or waiting time for rooms and personnel. The practitioners had to check the output of each schedule such as the assignments of operating rooms, anesthetists, nurses, and etc. to determine a preferred one. Similarly but differently, [102] executed a computer simulation program by considering "costs" of empty beds, hospital overflow, and loss of patients in different schedules.

Following the pioneer work in [101-102], many research efforts extended the scope of operating room scheduling by considering other practical objectives and/or constraints. For example, [103] applied a column-generation heuristic, a large-scale optimization algorithm, on a

weekly operating room and recovery room scheduling problem. [104] focused on patient waiting time in outpatient scheduling systems by using a simulation approach. [105] formulated a mixed-integer programming for surgeon and surgery scheduling enforcing constraints of surgeon health such as maximum number of time blocks per day, maximum number of consecutive working days, and etc. [106] took into consideration staff preferences and affinities, and listed affinity maximization as one of the objective functions.

In recent years, the research of operating room scheduling has been able to tackle uncertain factors because of the advances of solution techniques. For example, [107] combined Monte Carlo simulation and mixed-integer programming to solve operating room planning problems with random emergency surgery arrivals. [108] applied robust optimization with uncertain surgery duration for operating room scheduling without assuming any probability information. Also considering the randomness in duration of surgical procedures, [109] proposed both a stochastic programming model and a robust optimization model, and provided easy-to-implement heuristic solutions. We refer the readers to [110] for a more detailed review of such literature.

Outpatient surgery scheduling is a subtopic of surgery scheduling problem, which schedules mainly elective patients. In recent years, the aging population in the US increases demand for outpatient surgery services. By the year 2020, the forecasted growth depending on specialty is predicted to be 14% to 47% [112]. Improving efficiency and offering patients' greater convenience is critical. Ambulatory surgery centers (ASCs) have become more popular. In 2012, there were 5,357 ASCs, which received $3.6 billion in Medicare payments, according to MedPAC data[2]. In contrast to a large volume hospital, where the main objective is to maximize utilization, decrease downtime, and limit the overall cost, an ASC must flexibly adapt to a wide

variety of external realities, balancing service to providers against efficiency. In addition, in a large volume hospital, the operating suites run around the clock, often with dedicated ORs for emergent cases. Outpatient surgery centers, on the other hand, do not normally have emergent or urgent cases and patients do not require an overnight stay after the procedure. Yet in free-standing ASCs, increasing efficiency and reducing costs has become critical as competition has increased. The performance measures assign different priorities to some stakeholders over others. For the institution, increasing utilization and decreasing costs are important, as is supporting the service relationship with providers. For the patients, they desire no delay, short waiting times and high quality services. Physicians would like to have easy scheduling and more direct control. The objectives used in the scheduling system are adapted to the priorities in real system and the computational complexity.

There is an underlying need to develop methodologies and tools that will enable and assist managers in an ASC to schedule ORs more efficiently while maintaining needed flexibility. Existing scheduling system in most ASCs does not consider the stochastic behavior of operation duration which has a tremendous impact on the scheduling system overtime and idle time. And additional constraints can only be manually managed through experience and trial and error. Our study will help physicians and owners of ASCs to quickly identify problems in surgery scheduling, and thus recognize and capitalize on opportunities of improvement. Current OR scheduling systems in ASC will be enhanced by systemic scheduling tools that recommend the optimized schedule plan based on the real time setting. The broader impact will be achieved by the reduction of idle and overtime of ORs, further reducing the overall cost. Improved efficiency will also increase the satisfaction level of both patients and physicians by reducing patient waiting time, improving the accuracy of scheduling time and enhancing relationships between

surgery team members. The scheduling tools and the corresponding results are not limited to ASCs. They can be extended or tailored to other healthcare facilities, e.g., large scale hospital scheduling problems. Our studies are presented as follows.

First, we analyzed the historical data from an Ambulatory Surgery Center, and constructed the lognormal distributions for surgery durations of different types.

Second, we proposed a day-ahead nurse staffing model for the surgery center considering the fixed surgery schedules and the affinities between nurses and surgeons. The solution was implemented in AIMMS (a commercial tool for optimization modeling).

Finally, we formulated a multi-objective stochastic programming model for weekly operating room scheduling. We included the obtained lognormal distributions, and took into consideration the affinities between team members as well as the efficiencies of a nurse assistant in various surgeries. The solution was also implemented in AIMMS.

4.2 Estimation of Operation Time

4.2.1 Data Collection

For the purposes of this study, existing de-identified administrative data from university affiliated surgery center is used. The dataset initially includes time-related surgery information and clinical information (with prior recoding of patients and providers) from all surgical procedures and endoscopies for a 3-year period from 2011 to 2014. The total sample size is about 9000 with 580 procedures (identified by CPT procedure code). The time variables related to operations include patient pre-operation duration; time when a patient enters/leaves OR; time that a procedure starts; time that a patient arrives Recover phase I (PACU) as well as the time that a patient gets discharged (Recovery phase II). Other surgery information includes procedure code, the code numbers of the physicians and staff (de-identified names), the scheduled

procedure duration, and patient information limited to gender, age, American Society of Anesthesiologists (ASA) Score.

The scope of our study is to estimate operation duration and to produce scheduling strategy. The focus is on the time of ORs being occupied. Thus, operation time in this study is defined as the time between patient entering and leaving an operating room. Since different types of operations vary dramatically, time duration estimation is based on specialties.

4.2.2 Distribution Fitting

To estimate the distribution of the operation time, we started by grouping similar procedures. Since the 5-digit CPT coding system reflects the hierarchy of CPT codes, we checked the codes with the same first 4 digits. Some procedures are bundled to be more comprehensive. For example, code 11400, 11401, 11402, 11404, and 11406 are bundled because all of them are for excision of benign lesions of skin, with the only difference in sizes of excised diameter. However, some procedures are different even with the same first 4 digits. For instance, code 1582* reflects all sorts of "other facial procedures" and code 15820, 15822, 15823 are for blepharoplasty, but code 15824, 15825, and 15828 are coded as rhytidectomy. On the other hand, to show statistical significance in grouping, we also conducted k-sample Anderson-Darling test on the group candidates to test the hypothesis that k groups of data were drawn from the same distribution. Eventually, we were able to identify 50 groups of procedures, each of which has at least 30 records. Note that the records in those 50 groups cover more than 80% of the procedures in the original data.

To fit statistical distributions for those frequent operation durations, we tested the three most common time duration distributions: Gaussian distribution, Weibull distribution and lognormal distribution. We tested all 50 major procedures in the surgery center based on

maximal likelihood information. Most procedures (38 out of 50) show lognormal distribution as the best fit. We illustrate one example of distribution fitting comparison across different distributions. Procedure extracapsular cataract removal with insertion (CPT code is 66984) is the most frequent operation in the surgery center (2500 in total). In Figure 11(a-c), Weibull, Gaussian, and lognormal distribution are fitted to compare with the empirical distribution on the histogram plots respectively. As shown in the results, lognormal distribution is the closest to the empirical density. Since lognormal distribution is always close to the distribution with maximal likelihood, we use lognormal distribution for all operation durations to give consistent estimates. In [105], lognormal distribution is also applied to estimate surgery procedure durations.

In order to meet various and flexible expectations of scheduling results, we also provide 6 operation duration percentile values (50%, 60%, 70%, 80%, 90%, 95%) for each major procedure based on the obtained lognormal distributions. The larger percentile, the longer duration will be scheduled. For example, 50% of the procedures are estimated to be complete within 55 minutes; 60% are completed within 61 minutes; 70% within 67 minutes, and so on. When a larger percentile value is chosen, the likelihood of finishing operation on time is higher. On the other hand, choosing smaller percentile values will have more operations scheduled. Therefore, to choose a percentile value is a balance between efficiency and accuracy, and the final decision is on physicians and schedulers/center managers. To implement distribution fitting, we applied R package [111]. Figure 12 illustrates the information of the specialty groups (first 6 records are shown). The first column is the total number of records for that specialty; the second and third columns are the sample mean and median times (in minutes), respectively. Column 4 through column 7 are the parameters (mu and sigma) and their standard deviations of the fitted

lognormal distribution. The estimated mean value and 6 percentile values are shown in other columns.

Note that procedures (387 types in total) with fewer than 30 records could hardly reach any distribution. Min/max/mean/median values are provided as references to help schedule the operations, shown in Figure 13.

Based on the two tables above, we would have a better understanding in terms of surgery durations of historical procedures. Therefore, the scheduler could use the two tables as a reference for scheduling the major operations.

4.3 Daily Staffing Scheduling

In this subsection, we create a practical scheduling tool to solve daily staffing and allocation problems. By giving a set of operations with scheduled time duration, available ORs, and nurses, the proposed model can produce the room assignments and the optimal pairing teams by taking into consideration the service relationship with providers. The nomenclature used in this subsection is listed in Table 7.

The objective of the model is to maximize the total service efficiency, defined into two parts (7) and (8). (7) maximizes the total efficiency of nurse skill level of assisting operations. (8) maximizes the efficiency of surgery team (nurses and surgeons who prefer to work together have higher score).

$$max \sum_i \sum_n \sum_t E_{in} y_{inh} \tag{7}$$

$$max \sum_i \sum_n \sum_t \left(\sum_j C_{jn} B_{ij} \right) y_{inh} \tag{8}$$

The main constraints of the model are listed as following: A nurse must be available if she/he is assigned, defined by (9); A nurse can only be scheduled in one operation at a time, defined by (10); An operation can only be scheduled in consecutive slots, represented by (11)-

(13); Enough number of nurses for an operation (in (14)); Total working hour for nurses with the balance of under and over working time, represented by (15) -(17); Whether a nurse is scheduled or not (in (18)).

$$\sum_i z_{inh} \leq Q_{nh} X_{ih}, \forall i, n, h \tag{9}$$

$$\sum_i z_{inh} \leq 1, \forall n, h \tag{10}$$

$$\sum_n \sum_h z_{inh} = \sum_h X_{ih} \cdot M_i \,, \forall i \tag{11}$$

$$y_{inh} \geq z_{inh} - z_{in(h-1)}, \forall i, n, h = 2, 3, \dots, H \tag{12}$$

$$y_{inh} = z_{inh}, h = 1 \tag{13}$$

$$\sum_n \sum_h y_{inh} = M_i, \forall i \tag{14}$$

$$h_n F_n^{min} - w_n^{un} \leq \sum_i \sum_h z_{inh} \leq h_n F_n^{max} + w_n^{over}, \forall n \tag{15}$$

$$w_n^{un}, w_n^{over} \geq 0, \forall n \tag{16}$$

$$\sum_n (w_n^{un} + w_n^{over}) \leq 10 \tag{17}$$

$$h_n \geq y_{inh}, \forall i, n, h \tag{18}$$

We implemented the scheduling model in AIMMS with UI pages that schedulers could input available information. The panels are shown in Figure 14. The dark background panels are the input panels. The one with white background and a "solving" button is an output panel which shows the optimal staffing results. The left top corner panel shows the available time slots of each nurse (input values for parameter Q_{nh}) and scheduled operations (for parameter X_{ih}). We assume one slot equals to one hour initially, and it can be changed to other time unit, saying 30 minutes or 15 minutes. The shorter time for each slot, the more precise the results would be, but it shouldn't be too small to generate results in a reasonable time. In the left bottom corner, the upper panel shows the serving efficiency matrix between nurses and surgeons (C_{jn}). The value of C_{jn} is between 0 and 1, which needs to be given by the surgery center manager. The surgeon and

the values of nurses B_{ij} and $C_{jn}$are given for each operation in the lower panel. The default minimum and maximum working hours are given for each nurse (F_n^{min}, F_n^{max}). The results will be shown in right bottom panel.

4.4 Week-Ahead Stochastic Programming Scheduling

In this subsection, we propose a mathematical programming model to generate optimal schedules of surgeries and nurses, aiming at improving service efficiency as well as reducing the under/over working hours. The problem is formulated as a stochastic programming model due to the presence of uncertain operation durations. The nomenclature used in this subsection is listed in Table 8, and we try to accommodate as many constraints as possible.

$$max \sum_i \sum_n \sum_t E_{in} y_{int} \tag{19}$$

$$max \sum_i \sum_n \sum_t \left(\sum_j C_{jn} B_{ij}\right) y_{int} \tag{20}$$

$$min \sum_s \sum_j \sum_t \left(z_{jts}^{over}\right) + \sum_s \sum_n \sum_t (w_{nts}^{over}) \tag{21}$$

$$min \sum_i u_i \tag{22}$$

$$min \sum_i v_i \tag{23}$$

Multiple objectives are defined in (19)-(23). It is widely acknowledged that there generally more than one goal in operating room scheduling optimization. For example, (19)-(20) are formulated to maximize the efficiencies that nurses can assist surgeons for various surgeries; (21) tries to minimize the under or over working hours for both surgeons and nurses. In addition, "the loss of operations" is allowed in (22) to ensure the feasibility of the proposed model, but will be minimized by associating a very large penalty to any unscheduled operation. Finally, patient waiting days, often used in literature as a measurement of healthcare delivery quality, is captured in (23). There are different approaches to deal with multiple objectives in optimization, for example, goal programming, linear scalarization, etc. Linear scalarization is able to find

Pareto-optimal solutions, but requires to association of different coefficients to formulate a single objective function (i.e., the sum of weighted objectives). But it is very debatable on how to choose or assign different weights to various objectives especially in healthcare related research. Therefore, we use goal programming as the solution approach in this section, and the details will be presented later. The complexity of the stochastic scheduling problem stems from the following constraints.

$$\sum_k \sum_t x_{ikt} = 1 - u_i, \forall i \tag{24}$$

$$x_{ikt} \leq R_{kt} S_{ik} \sum_j \left(P_{jt} \cdot B_{ij}\right), \forall i, k, t \tag{25}$$

$$y_{int} \leq Q_{nt} \sum_k x_{ikt}, \forall i, n, t \tag{26}$$

$$(1 - u_i) T_i^{min} \leq \sum_k \sum_t (x_{ikt} \cdot t) \leq (1 - u_i) T_i^{max}, \forall i \tag{27}$$

$$\sum_k \sum_t (x_{ikt} \cdot t) - T_i^{min} \leq v_i, \forall i \tag{28}$$

$$u_i \left(T - T_i^{min}\right) \leq v_i, \forall i \tag{29}$$

Constraint (24) ensures that a surgery must be scheduled to avoid a large penalty. Constraint (25) states that a surgery can be scheduled to a room only if the room is available. Similarly, the availability of a nurse is enforced in (26). Constraints (27)-(29) restrict the earliest and due date of a scheduled surgery.

$$\sum_t \sum_n y_{int} = M_i, \forall i \tag{30}$$

$$\sum_k \sum_i x_{ikt} \cdot B_{ij} \cdot D_{is} \leq h_{jt}^{sur} G_{jt}^{max} + z_{jts}^{over}, \forall j, t, s \tag{31}$$

$$z_{jts}^{over} \geq 0, \forall j, t, s \tag{32}$$

$$\sum_i y_{int} D_{is} \leq h_{nt}^{nur} F_{nt}^{max} + w_{nts}^{over}, \forall n, t, s \tag{33}$$

$$w_{nts}^{over} \geq 0, \forall n, t, s \tag{34}$$

Constraint (30) guarantees sufficient number of nurses is assigned for each surgery. The over working hours of surgeons and nurses are modeled in (31)-(34).

$$h_{jt}^{sur} \geq x_{ikt} \cdot B_{ij}, \forall i, j, k, t \quad (35)$$

$$h_{nt}^{nur} \geq y_{int}, \forall i, n, t \quad (36)$$

$$\sum_i x_{ikt} \left(\sum_s \pi_s D_{is}\right) \leq 8, \forall k, t \quad (37)$$

Note that (35) and (36) indicate that there will be no over-time penalty if a surgeon or a nurse is not scheduled for some day. Finally, constraint (37) limits the number of available hours for each operating room.

As mentioned earlier, goal programming is used to solve the above multi-objective stochastic programming. The idea is to solve the stochastic programming with one objective first, and then convert this objective to a constraint whose limit is the optimal objective value. The process is repeated until all objective functions in the original problem are included. Let SP stand for the feasibility region of the proposed stochastic programming, i.e., constraint (24)-(37), and let O^* stand for the optimal objective value when (*) is the objective. The whole procedure is shown in Figure 15. Note that one disadvantage of the goal programming is the objective functions included in early steps will always take priority. Thus at some step, the optimal objective could be relaxed when this objective is converted to a constraint. This can enlarge the feasibility regions for the optimization problems in the followed steps. For example, a coefficient 1.1 is added in the final step of Figure 15.

We used the data from the surgery center to test our proposed model and algorithm. The surgery data from the weeks of Jul-08-2013, Jan-10-2014, and May-12-2014 were randomly chosen for our numerical study. There are 68, 63, 62 surgeries in these three weeks, respectively. There are four operating rooms available for around 20 surgeons each week. A number of 10 scenarios are sampled for each week, which follows the lognormal distributions presented in section 4.2. The probability of each scenario is 10%. In Figure 16, a few case samples from week

2 are shown. The first 3 column represent Case number, Surgeon number (de-identified), and CPT code, respectively. Column S1 to S10 are the 10 scenarios generated based on the procedure distribution (distribution table shown in Figure 12). Note that almost 90% of those procedures have distributions, thus the median value will be applied across all scenarios if a procedure duration distribution is not available for a surgery (median values shown in Figure 13). In Figure 16, case 44742, case 44977 and case 44990 are infrequent procedures, median values of the durations being applied for sampling. The efficiency of a nurse assisting surgeons for different surgeries is missing, and is randomly generated with a [0.5, 1] uniform distribution. We assume there are eight nurses working in the surgery center. The goal programming is implemented in Aimms 4.1. The stopping criteria is 1% for mixed-integer programming gaps and 10 minutes for solution time.

The computational results are listed in Table 9, Table 10 and Table 11 for week 1, 2, and 3 respectively. The achieved gap guarantees the quality of solutions in each step. Also based on our testing, the first two stochastic programming, i.e., minimizing "loss of surgeries" and "surgeon/nurse over-working hours", can be solved in a very short time. The computation of the last step regarding "affinities and efficiencies" is stable and not time-consuming. The most challenging part is to minimize patient waiting days [104]. The overall performance of this goal programming validates that the proposed model could be used as a decision-making tool to schedule week-ahead operations and nurses by considering multiple objectives.

4.5 Conclusion

Our study provides multi-level scheduling solutions to help improve efficiency and service of a surgery center without increasing the complexity of the system. These scheduling methods can be implemented individually or combined together as recommended options for

schedulers with simple query and input. For example, operation duration distribution estimations and day-ahead scheduling, presented in section 4.2 and 4.2 respectively, could be implemented directly as a simple two-part scheduling recommendation tool to a surgery center. Consequently, percentile duration estimations could be used to better distribute the optimal operation time to physicians, and then day-ahead information can be plugged in to produce immediate daily optimized rostering solution.

A promising direction for future study is to analyze scenario reduction for week-ahead surgery planning. To compromise with computation, the number of sampled scenarios is usually limited. But the more scenarios we have, the more precise the stochastic programming is. Therefore, it is worth to try different scenarios reduction techniques such as clustering based on deterministic objective functions, the similarity of the first-stage decision variables, and etc. Another avenue we could explore is chance-constrained programming which does not require any sampling, and can provide a confidence level for each constraint.

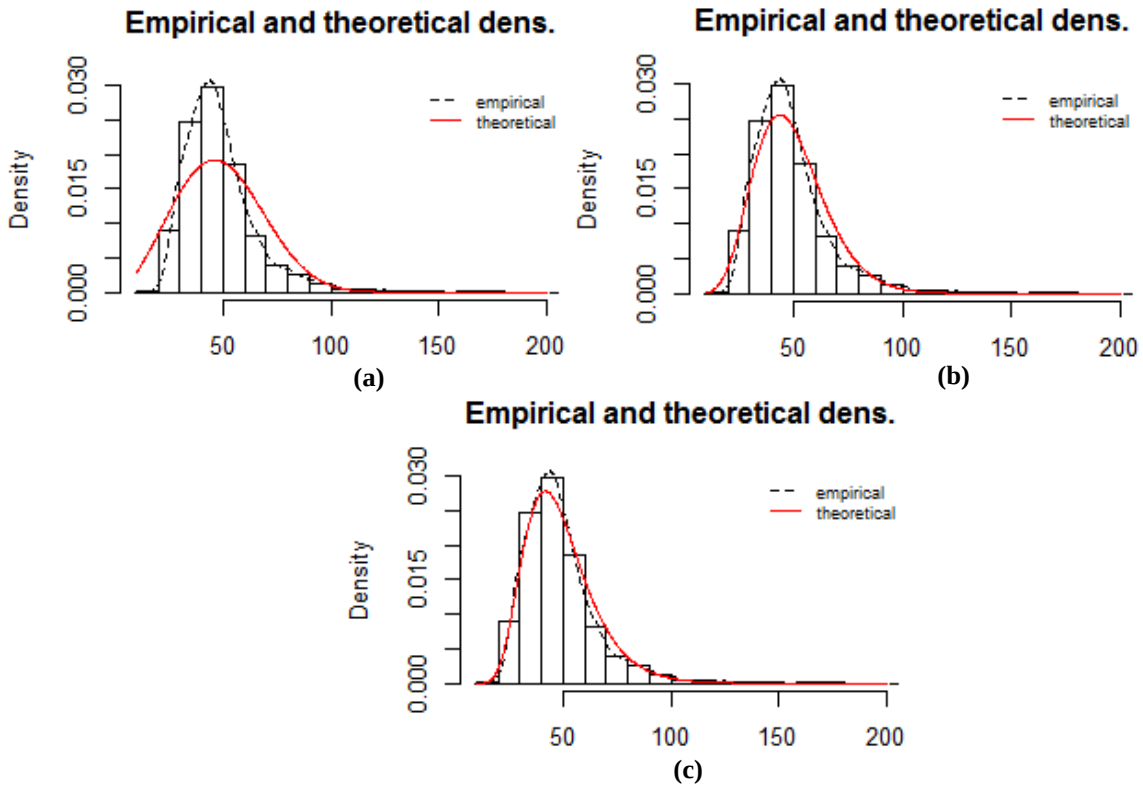

Figure 11 Comparison of empirical and theoretical densities. ((a)Weibull distribution, (b) Gamma distribution, and (c) Lognormal distribution)

	num	s_mean	s_med	mu	mu_sd	sig	sig_sd	e_mean	e_med	e_60	e_70	e_80	e_90	e_95	CPT
1	38	59	54	4.01	0.06	0.37	0.04	59	55	61	67	76	89	102	11400 11401 11402 11404 11406
2	36	62	55	4.05	0.06	0.36	0.04	61	58	63	70	78	92	105	11420 11421 11422 11423 11424 11426
3	32	69	64	4.12	0.08	0.48	0.06	69	61	69	79	92	113	134	11440 11441 11442 11443 11444 11446
4	73	117	110	4.7	0.05	0.39	0.03	118	109	121	134	151	179	206	11641 11642 11643 11644 11646
5	176	116	90	4.61	0.04	0.52	0.03	115	101	115	132	156	195	236	15820 15822 15823
6	35	305	296	5.66	0.06	0.36	0.04	307	288	315	347	389	455	517	15824 15825 15828

Figure 12 Lognormal distribution information of frequent procedures. (first 6 records)

	cpt	num	mean	median	min	max
1	10121	1	34	34	34	34
2	10140	2	50	50	50	50
3	10160	7	63	54	38	103
4	10180	3	48	43	43	57
5	11000	1	56	56	56	56
6	11010	7	77	76	37	121

Figure 13 Descriptive statistics of infrequent procedures. (first 6 records)

Table 7 Nomenclature of daily staffing scheduling model

i	Operation, $i = 1, 2, \dots, I$
j	Surgeon, $j = 1, 2, \dots, J$
n	Nurse, $n = 1, 2, \dots, N$
h	Time slot, $h = 1, 2, \dots, H$
B_{ij}	Surgeon j with Operation i
Q_{nh}	1 if Nurse n is Available on Time slot h; 0 Otherwise
X_{ih}	1 if Operation i is scheduled in Time slot h; 0 Otherwise
F_n^{min}, F_n^{max}	Min/Max Working Hours of Nurse n
E_{in}	Efficiency Nurse n can Assist Operation i
C_{jn}	Efficiency between i Surgeon j and Nurse n
M_i	Number of Nurses Required for Operation i
y_{inh}	Binary Variable, 1 if Nurse n starts from Slot h for Operation i; 0 Otherwise
z_{inh}	Binary Variable, 1 if Nurse n works in Slot h for Operation i; 0 Otherwise
w_n^{un}, w_n^{over}	Continuous Variable, Under/Over Working Time of Nurse n
h_n	Binary Variable, 1 if Nurse n is assigned to assist operations

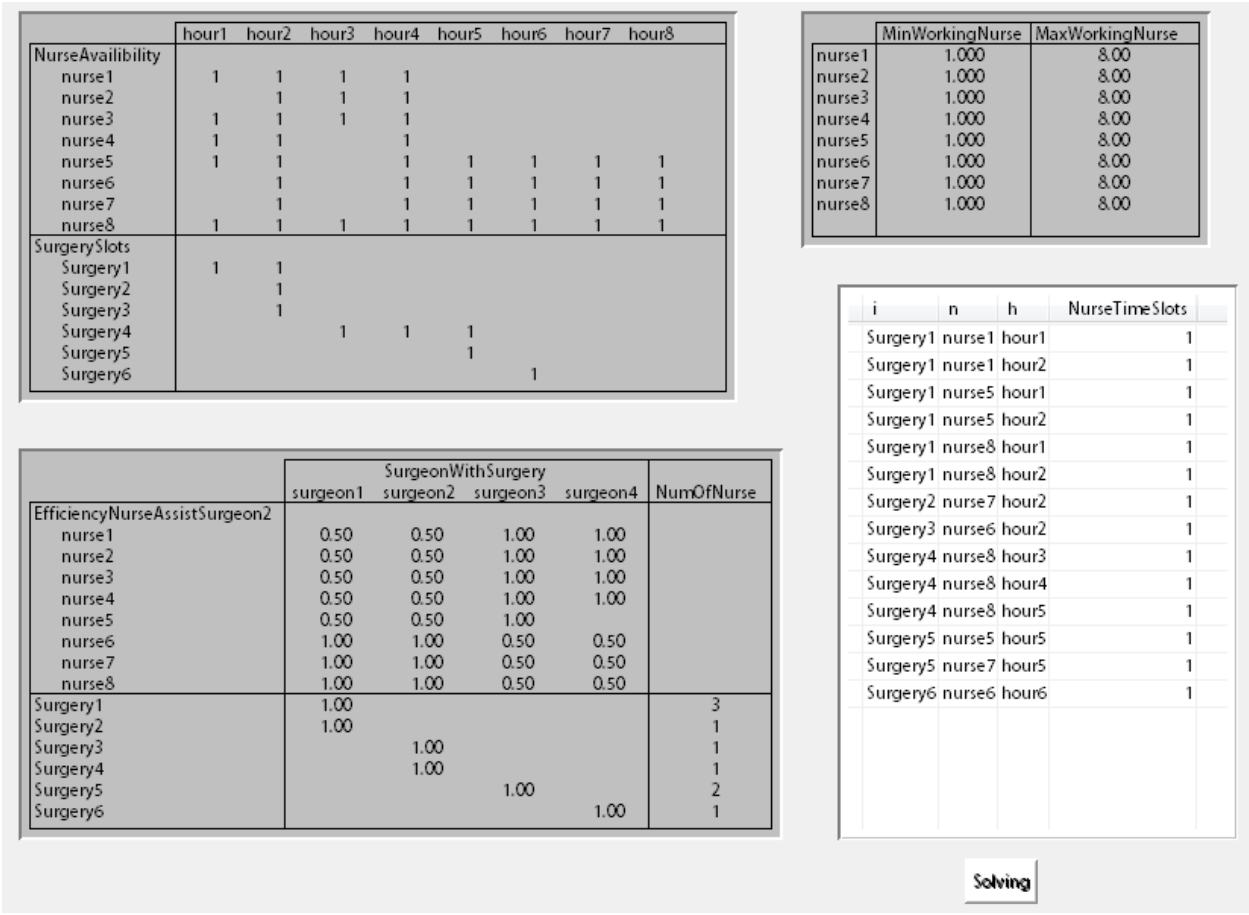

Figure 14 Graphical user interface of day-ahead scheduling tool

Table 8 Nomenclature of week-ahead stochastic programming scheduling model

i	Operation, $i = 1, 2, \dots, I$
j	Surgeon, $j = 1, 2, \dots, J$
n	Nurse, $n = 1, 2, \dots, N$
k	Room, $k = 1, 2, \dots, K$
t	Day, $j = 1, 2, \dots, T$
P_{jt}	1 if Surgeon j is Available on Day t; 0 Otherwise
Q_{nt}	1 if Nurse n is Available on Day t; 0 Otherwise
R_{kt}	1 if Room k is Available on Day t; 0 Otherwise
T_i^{min}, T_i^{max}	Earliest/Due Date for Operation i
G_j^{max}	Max Working Hours of Surgeon j on Day t
F_n^{max}	Max Working Hours of Nurse n on Day t
E_{in}	Efficiency Nurse n can Assist Operation i
C_{jn}	Efficiency between i Surgeon j and Nurse n
B_{ij}	Surgeon j with Operation i
S_{ik}	Operation i can be operated in Room k
M_i	Number of Nurses Required for Operation i
D_{is}	Normal Opertion Time for Operation i of scenario s
π_s	Probability of scenario s
x_{ikt}	Binary Variable, 1 if Operation i is assigned in Room k on Day t; 0 Otherwise
y_{int}	Binary Variable, 1 if Nurse is assigned on Day t for Operation i; 0 Otherwise
z_{jts}^{over}	Continuous Variable, Over Working Time of Surgeon j on Day t in scenario s
w_{nts}^{over}	Continuous Variable, Over Working Time of Nurse n on Day t in scenario s
v_i	Integer Variable, Patient i Waiting Days
u_i	Binary Variable, 1 if Operation i is not assigned successfully
h_{jt}^{sur}	Binary Variable, 1 if Surgeon j is assigned operations on day t
h_{nt}^{n}	Binary Variable, 1 if Nurse n is assigned to assist operations on day t

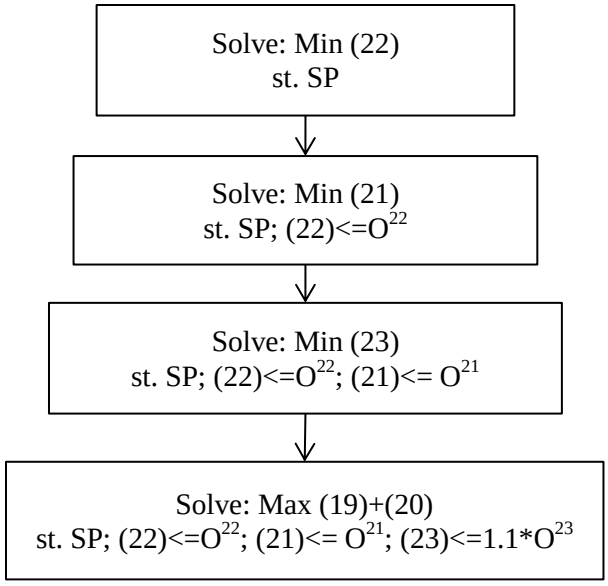

Figure 15 Goal programming for proposed model

Case #	Surgeon #	CPT code	Table	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
42939	P27098	29805	1	51	143	123	114	130	90	194	158	82	165
43886	P39576	67904	1	67	57	68	92	138	145	134	77	73	66
44404	S00638	58558	1	57	43	55	47	41	52	98	61	90	79
44742	D00637	19371	2	225	225	225	225	225	225	225	225	225	225
44831	P39576	67904	1	55	130	81	127	93	165	146	91	153	49
44977	S00623	31231	2	99	99	99	99	99	99	99	99	99	99
44985	P99900	19318	1	168	87	225	91	109	152	141	352	114	145
44990	P39576	14060	2	74	74	74	74	74	74	74	74	74	74
45095	P65899	15836	1	229	275	294	425	143	255	300	176	156	145
45122	P39576	67917	1	62	130	59	47	49	83	149	79	87	96
45204	P00112	66984	1	37	77	71	29	29	24	44	45	66	54
45205	P00112	66984	1	28	79	47	58	51	26	59	30	66	40
45207	P00112	66984	1	60	70	43	34	29	55	42	40	70	43
45208	P00112	66984	1	58	56	49	33	78	88	36	40	46	36
45209	P00112	66984	1	21	36	80	49	40	46	64	48	45	37
45210	P00112	66984	1	58	33	57	49	113	46	77	46	55	48
45320	I65830	30520	1	169	85	141	143	117	164	214	167	143	109

Figure 16 Scenarios of case samples from week 2

Table 9 Computational results of goal programming for week 1

Step	Objective Value	Solution Time (seconds)	Achieved MIP Gap
1	0	0.19	0%
2	0.9	1.03	0%
3	43	603.4	2.49%
4	93.7	58.33	0.75%

Table 10 Computational results of goal programming for week 2

Step	Objective Value	Solution Time (seconds)	Achieved MIP Gap
1	0	0.13	0%
2	0	0.55	0%
3	34	22.46	0%
4	89.7	32	0.67%

Table 11 Computational results of goal programming for week 3

Step	Objective Value	Solution Time (seconds)	Achieved MIP Gap
1	0	0.12	0%
2	0	0.61	0%
3	40	7.36	0%
4	84.2	5.85	0.83%

CHAPTER 5: CONCLUSION

Our studies presented in this dissertation have advanced the knowledge and techniques in healthcare delivery including hospital readmissions, nonlinear associations, and operating room scheduling. On one hand, several traditional and newly developed statistical techniques and data mining algorithms were applied to different medical data and patient groups, either confirming existing knowledge established by other methodologies or discovering new information in healthcare delivery field that could benefit both healthcare providers and patients. On the other hand, a set of scheduling tools were developed to improve efficiencies of healthcare planning processes and experiences of healthcare receivers. Both statistical and operation research methods were conducted to provide parameter estimations/variations, and to generate optimal solutions.

First, we deepened the understanding of risk factors associated with unplanned readmissions in pre-specified disease cohorts. We chose a 30-day readmission rate, which is well-established and widely-accepted in healthcare industry. We successfully identified factors associated with the patient, disease severity, and hospital stay. Different mixes of risk factors are generated for five chronic diseases by the two methods in our study, the LR model and the proportional hazards model. We furthered our understanding on those factors by analyzing specific factors for each given disease. In most cases, the significant factors were consistent across all of those diseases and can be explained. However, there are a few findings that are difficult to argue, especially the limitation of our dataset. In future studies, other clinical factors

related to patients should be included into predictive models. This might require data or records beyond current administrative claims data, but it will definitely extend the boundary of risk factors in readmission study. Different data mining algorithms can be applied to explore more complicated data structures and identify the associations between different factors in a specific disease. Finally, to study a large number of patients with different disease combinations could lead to a closer look of the risk factors in case of intense interactions.

Second, we studied patients underwent procedures who have also been readmitted within 30 days. Our study illustrated that traditional and advanced analytical methods fail to draw conclusions on the whole population with low predictive accuracy. However, on the subpopulation level, conditional inference tree analysis identified extremely high risk patient groups. It is the first time conditional tree method being used on ACS-NSQIP data. The results show that short postoperative length of stay could be a main reason for certain patients to get readmitted, patients with organ space surgical site infection and patients who return to OR during hospital stay. The patients with those conditions should be carefully examined before discharge and performed with close follow up exams. This study is also a retrospective analysis and restricted to colorectal surgery patients while patients who underwent other procedures could have different postoperative risk factors. Our future work is to help conduct prospective study to investigate more hospital risk factors within the hospital system. When a patient is identified as high risk patient, an inter-disciplinary research team including physicians, nurses, pharmacists, and social workers would be brought together to explore and learn how to decrease the risk of readmission by looking at each case individually.

Third, a novel graphical model, the sparse tree-embedded graphical model (STGM), is proposed to detect clinical associations. The proposed STGM can uncover both linear and

nonlinear relationships from a large number of variables. The basic idea of our STGM is integrating regression-based methods with decision tree learning. We further proposed an efficient regression-based algorithm for learning the STGM from data. We conducted simulation studies that demonstrated superiority of the STGM over other network learning methods, and applied the STGM on patients with Type II diabetes that demonstrated its efficacy on discovering interesting nonlinear relationships in practice. Future research directions include the investigation of how to extend STGM to discover more kinds of nonlinear relationships, how to provide insights for guiding better allocation of intervention resources, and how to provide baseline treatment guidelines for caregivers and clinicians by combining diagnosis and treatment information together. Also, in additional to diagnosis code, significant patients characteristics related to disease could be put into consideration, like age, gender, social-economic status, etc.

Finally, we proposed and constructed a multi-level scheduling solution to help improve efficiency and service of surgery center without increasing the complexity of the system. It involves a study of surgery duration distributions on more than 50 surgery types that present frequently in an ambulance surgery center. Those distributions appear to be very attractive to the planners, as they can schedule the surgeries based on statistical distributions instead of in an Ad-hoc manner. Moreover, a day-ahead nurse staffing tool is proposed and implemented to generate optimal nurse scheduling with the consideration of the efficiencies that a nurse can assist a surgery. This model stems from the current practice that a surgeon usually work better with some of the nurses. The time duration estimates and staffing tools can be implemented individually or combined together as recommended options for schedulers with simple query and input. Consequently, percentile duration estimations could be used to better distribute the optimal operation time to physicians, then day-ahead information can be plugged in to produce

immediate daily optimized rostering solution. Last but not the least, we proposed a week-ahead scheduling model, which is formulated as a stochastic programming, to tackle the uncertainties the decision maker faces ahead of time. This stochastic programming is built upon the distributions we obtained at the beginning of this study. The recommended solution might not be feasible in reality due to the flexible schedules of physicians and patients. However, our model provides a theoretical guideline and a long-term planning reference for operating room and nurse scheduling. Our future work is to test and improve the models in the surgery centers and other big volume hospitals. The models would be modified based on various health system settings such as the number of ORs, nurses, and additional healthcare providers and social workers. Accordingly, tailored mathematical programming techniques will be adopted to overcome computational obstacles if necessary.

REFERENCES

[1] 2013 National Healthcare Quality Report (NHQR). Available from: http://www.ahrq.gov/research/findings/nhqrdr/nhqr13/2013nhqr.pdf.

[2] Report to the Congress: Medicare Payment Policy, March 2014. Medicare Payment Advisory Commission (MedPAC) report. Available from: http://www.medpac.gov/documents/reports/mar14_entirereport.pdf?sfvrsn=0

[3] Jencks, S. F., Williams, M. V, Coleman, E., 2009. Rehospitalizations among patients in the Medicare fee-for-service program. The New England Journal of Medicine, 360(14), 1418-1428.

[4] The revolving door: A report on U.S. hospital readmissions. 2013 Robert Wood Johnson Foundation. Available from: http://www.rwjf.org/content/dam/farm/reports/reports/2013/rwjf404178.

[5] Garner, C. B., 2014. Medicare: The Gift that Keeps on Giving. Corporate Compliance Insights, September.

[6] Rico, F., Liu, Y., Martinez, D. A., Huang, S., Zayas-Castro, J. L., 2015.Preventable Readmission Risk Factors for Patients with Chronic Conditions. To Appear in Journal for Healthcare Quality.

[7] Stone, J., Hoffman, G. J., 2010. Medicare Hospital Readmissions : Issues, Policy Options and PPACA. Report for Congress

[8] Hamner, J. B., Ellison, K. J., 2005. Predictors of hospital readmission after discharge in patients with congestive heart failure. Heart & Lung: The Journal of Acute and Critical Care, 34(4), 231-239.

[9] Keenan, P. S., Normand, S.-L. T., Lin, Z., Drye, E. E., Bhat, K. R., Ross, J. S., Krumholz, H. M., 2008. An administrative claims measure suitable for profiling hospital performance on the basis of 30-day all-cause readmission rates among patients with heart failure. Circulation. Cardiovascular Quality and Outcomes, 1(1), 29-37.

[10] Kosiborod, M., Smith, G. L., Radford, M. J., Foody, J. M., Krumholz, H. M., 2003. The prognostic importance of anemia in patients with heart failure. The American Journal of Medicine, 114(2).

[11] Lindenauer, P. K., Bernheim, S. M., Grady, J. N., Lin, Z., Wang, Y., Wang, Y., Krumholz, H. M., 2010. The performance of US hospitals as reflected in risk-standardized 30-day mortality and readmission rates for medicare beneficiaries with pneumonia. Journal of Hospital Medicine : An Official Publication of the Society of Hospital Medicine, 5(6), 12-18.

[12] Greenblatt, D. Y., Weber, S. M., O'Connor, E. S., LoConte, N. K., Liou, J. I., Smith, M., 2010. Readmission after colectomy for cancer predicts one-year mortality. Annals of Surgery, 251(4), 659-669

[13] Raval, A.D., Zhou, S., Wei, W., Bhattacharjee, S., Miao, R., Sambamoorthi, U., 2015. 30-Day Readmission Among Elderly Medicare Beneficiaries with Type 2 Diabetes. Popul Health Manag. Jan 21.

[14] Curtis, J. P., Schreiner, G., Wang, Y., Chen, J., Spertus, J. a, Rumsfeld, J. S., Krumholz, H. M., 2009. All-cause readmission and repeat revascularization after percutaneous coronary intervention in a cohort of medicare patients. Journal of the American College of Cardiology, 54(10), 903-907.

[15] Frei-jones, M. J., Field, J. J., 2009. Risk Factors for Hospital Readmission Within 30 Days : A New Quality Measure for Children With Sickle Cell Disease. Pediatr Blood Cancer, December 2008, 481-485.

[16] Allaudeen, N., Vidyarthi, A., Maselli, J., Auerbach, A., 2011. Redefining readmission risk factors for general medicine patients. Journal of Hospital Medicine : An Official Publication of the Society of Hospital Medicine, 6(2), 54-60.

[17] Bahadori, K., FitzGerald, J. M., Levy, R. D., Fera, T., Swiston, J., 2009. Risk factors and outcomes associated with chronic obstructive pulmonary disease exacerbations requiring hospitalization. Canadian Respiratory Journal : Journal of the Canadian Thoracic Society, 16(4), 43-49.

[18] Berman, K., Sweta, T., Forsell, K., Vuppalanch, R., Burton, J., Nguyen, J., Chalasani, N., 2011. Incidence and predictors of 30-day readmission among patients hospitalized for advanced liver disease. Clinical Gastroenterology and Hepatology, 9, 254-259.

[19] Callaly, T., Hyland, M., Trauer, T., Dodd, S., Berk, M., 2010. Readmission to an acute psychiatric unit within 28 days of discharge: identifying those at risk. Australian Health Review : A Publication of the Australian Hospital Association, 34(3), 282-285.

[20] Feudtner, C., Levin, J. E., Srivastava, R., Goodman, D. M., Slonim, A. D., Sharma, V., Hall, M., 2009. How well can hospital readmission be predicted in a cohort of hospitalized children? A retrospective, multicenter study. Pediatrics, 123(1), 286-293.

[21] Hartney, M., Liu, Y., Velanovich, V., Fabri, P., Marcet, J., Grieco, M., Huang, S., Zayas-Castro, J. L., 2014. Bounceback branchpoints: Using conditional inference trees to analyze readmissions. Surgery, 156 (4), 842-848.

[22] Nantsupawat, T., Limsuwat, C., Nugent, K., 2012. Factors affecting chronic obstructive pulmonary disease early rehospitalization. Chronic Respiratory Disease, 9(2), 93-98.

[23] Neupane, B., Walter, S. D., Krueger, P., Marrie, T., Loeb, M. 2010. Predictors of in hospital mortality and re-hospitalization in older adults with community-acquired pneumonia: a prospective cohort study. BMC Geriatrics, 10(1), 22.

[24] Whitlock, T. L., Repas, K., Tignor, A., Conwell, D., Singh, V., Banks, P. a, Wu, B. U., 2010. Early readmission in acute pancreatitis: incidence and risk factors. The American Journal of Gastroenterology, 105(11), 2492-2497.

[25] Belfort, M. a, Clark, S. L., Saade, G. R., Kleja, K., Dildy, G. a, Van Veen, T. R., Kofford, S., 2010. Hospital readmission after delivery: evidence for an increased incidence of nonurogenital infection in the immediate postpartum period. American Journal of Obstetrics and Gynecology, 202(1), 1-7.

[26] Khawaja, F. J., Shah, N. D., Lennon, R. J., Slusser, J. P., Alkatib, A. a, Rihal, C. S., Ting, H. H., 2012. Factors associated with 30-day readmission rates after percutaneous coronary intervention. Archives of Internal Medicine, 172(2), 112-117.

[27] Alkalay, A. L., Bresee, C. J., Simmons, C. F., 2010. Decreased neonatal jaundice readmission rate after implementing hyperbilirubinemia guidelines and universal screening for bilirubin. Clinical Pediatrics, 49(9), 830-833.

[28] Courtney, M., Edwards, H., Chang, A., Parker, A., Finlayson, K., Hamilton, K., 2009. Fewer emergency readmissions and better quality of life for older adults at risk of hospital readmission: a randomized controlled trial to determine the effectiveness of a 24-week exercise and telephone follow-up program. Journal of the American Geriatrics Society, 57(3), 395-402.

[29] Horwitz, L., Partovian, C., Lin, Z., Herrin, J., Grady, J., 2012. Hospital-Wide All-Cause Unplanned Readmission Measure final report.

[30] ICD-9-CM 6th ed., 2010. Publised by Jackson, Wyo.

[31] Liu, Y., Zayas-Castro, J. L., Fabri, P., Huang, S., 2014. Learning high-dimensional networks with nonlinear interactions by a novel tree-embedded graphical model. Pattern Recognition Letters, 49, 207-213.

[32] Charlson, M., Szatrowski, T. P., Peterson, Janey Gold, J., 1994. Validation of a combined comorbidity index. Journal of Clinical Epidemiology, 47(11), 1245-1251.

[33] Box-Steffensmeier, J. M., De Boef, S., 2006. Repeated events survival models: the conditional frailty model. Statistics in Medicine, 25(20), 3518-3533.

[34] Kassin, M.T., Owen, R.M., Perez, S.D., Leeds, I., Cox, J.C., Schnier, K., 2012. Risk factors for 30-day hospital readmission among general surgery patients. J Am Coll Surg, 215, 322-330.

[35] Ozturk, E., Kiran, R.P., Remzi, F., Fazio, V. W., 2009. Early readmission after ileoanal pouch surgery. Dis Colon Rectum, 52, 1848-1853.

[36] Drolet, S., Maclean, A. R., Myers, R. P., Shaheen, A. A. M., Dixon, E., Buie, W. D., 2010. Morbidity and mortality following colorectal surgery in patients with end-stage renal failure: a population-based study. Diseases of the Colon & Rectum, 53(11), 1508-1516.

[37] Longo, W. E., Virgo, K. S.,Johnson, F. E., Oprian , C. A., Vernava, A. M., Wade, T. P., Phelan, M. A., Henderson, W. G., Daley, J., Khuri, S. F., 2000. Risk factors for morbidity and mortality after colectomy for colon cancer. Diseases of the Colon & Rectum, 43(1): 83-91.

[38] Spanjersberg, W. R., Reurings, J.,Keus, F., Van Laarhoven , C., 2011. Fast track surgery versus conventional recovery strategies for colorectal surgery. Cochrane Database Syst Rev 2.

[39] Li, L. T., Mills, W. L., White, D. L., Li, A., Gutierrez , A. M., Berger, D. H., Naik, A. D. 2013. Causes and Prevalence of Unplanned Readmissions After Colorectal Surgery: A Systematic Review and MetaAnalysis. Journal of the American Geriatrics Society, 61(7), 1175-1181.

[40] Wick, E.C., Shore, A.D., Hirose, K., Ibrahim, A.M., Gearhart, S.L., Efron, J., 2011. Readmission rates and cost following colorectal surgery. Dis Colon Rectum, 54, 1475-1479.

[41] Schneider, E.B., Hyder, O., Brooke, B.S., Efron, J., Cameron, J.L., Edil, B.H., 2012. Patient readmission and mortality after colorectal surgery for colon cancer: impact of length of stay relative to other clinical factors. J Am Coll Surg, 214, 390-398.

[42] Adamina, M., Kehlet, H., Tomlinson , G. A., Senagore, A. J., Delaney, C. P., 2011. Enhanced recovery pathways optimize health outcomes and resource utilization: a meta-analysis of randomized controlled trials in colorectal surgery. Surgery, 149(6), 830-840.

[43] Aarts, M. A., Okrainec, A., Glicksman, A., Pearsall, E., Victor, J. C., McLeod, R. S., 2012. Adoption of enhanced recovery after surgery (ERAS) strategies for colorectal surgery at academic teaching hospitals and impact on total length of hospital stay. Surgical Endoscopy, 26(2), 442-450.

[44] Kehlet, H., Wilmore, D. W., 2008. Evidence-based surgical care and the evolution of fast-track surgery. Annals of surgery 248(2), 189-198.

[45] Kehlet, H. Slim, K., 2012. The future of fast track surgery. British Journal of Surgery 99(8), 1025.

[46] Fiore, J., Browning , L., Bialocerkowski , A., Gruen, R., Faragher, I., Denehy, L., 2012. Hospital discharge criteria following colorectal surgery: a systematic review. Colorectal Disease, 14(3), 270-281.

[47] Tsai, T. C., Joynt, K. E., Orav, E. J., Gawande, A. A., Jha, A. K., 2013. Variation in Surgical-Readmission Rates and Quality of Hospital Care. New England Journal of Medicine, 369(12), 1134-1142.

[48] Birkmeyer, J. D., T. A. Stukel, A. E. Siewers, P. P. Goodney, D. E. Wennberg and F. L. Lucas, 2003. Surgeon volume and operative mortality in the United States. New England Journal of Medicine, 349(22), 2117-2127.

[49] Krell, R. W., Girotti, M. E., Fritze, D., Campbell, D. A., Hendren, S., 2013. Hospital readmissions after colectomy: a population-based study. Journal of the American College of Surgeons, 217(6), 1070-1079.

[50] American College of Surgeons National Surgical Quality Improvement Program (ACS NSQIP): Program Overview, 2012. Available from: http://site.acsnsqip.org/wp-content/uploads/2012/11/NSQIP-Overview-10.12.pdf

[51] Yu,S., Yu, K., Tresp, V., Kriegel, H. P., Wu, M., 2006. Supervised Probabilistic Principal Component Analysis. Proceeedings of the 12th ACM SIGKDD international conference on Knowledge Discovery and Data Mining, 464-473.

[52] Yamamoto, H., Yamaji, H., Abe, Y., Harada, K., Waluyo, D., Fukusaki, E., Kondo, A., Ohno, H., Fukuda, H., 2009. Dimensionality reduction for metabolome data using PCA, PLS, OPLS, and RFDA with differential penalties to latent variables. Chemometrics and Intelligent Laboratory Systems, 98(2), 136-142.

[53] Statnikov, A., Wang L., Aliferis, C. F., 2008. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. BMC Bioinformatics, 9, 319.

[54] Hothorna, T., Hornika, K., Zeileis, A., 2006. Unbiased recursive partitioning: A conditional inference framework. Journal of Computational and Graphical Statistics, 15(3), 651-674.

[55] Caliendo, M., Kopeinig, S., 2008. Some practical guidance for the implementation of propensity score matching. Journal of economic surveys, 22(1), 31-72.

[56] Friedman, N., Linial, M., Nachman, I., Pe'er, D, 2000. Using Bayesian networks to analyze expression data. Journal of Computational Biology, 7(3-4), 601–620.

[57] Huang, S., Li, J., Sun, L., Wu, T., Chen, K., Fleisher, A., Reiman, E., Ye, J., 2010. Learning brain connectivity of Alzheimer's disease by sparse inverse covariance estimation. NeuroImage, 50, 935-949.

[58] Dempster, A. P., 1972. Covariance selection, Biometrics, 28(1), 157-175.

[59] Drton, M., Perlman, M.D., 2004. Model selection for Gaussian concentration graphs. Biometrika, 91(3), 591-602.

[60] Drton, M.; Perlman, M.D., 2007. Multiple testing and error control in Gaussian graphical model selection. Statistical Science, 22(3), 430-449.

[61] Spirtes, P., Glymour, C., Scheines, R., 2000. Causation, Prediction, and Search, 2nd edition. MIT Press, MA.

[62] Whittaker, J., 1990. Graphical Models in Applied Multivariate Statistics. Wiley, New York.

[63] Friedman J, Hastie T, Hofling H, Tibshirani R., 2007. Pathwise Coordinate Optimization. The Annals of Applied Statistics, 1(2), 302-332.

[64] Ravikumar, P., Raskutti, G., Wainwright, M. J., Yu, B., 2008. Model selection in Gaussian graphical models: high-dimensional consistency of l1-regularized MLE. Advances in Neural Information Processing Systems (NIPS), 21.

[65] Yuan, M., Lin, Y., 2007. Model selection and estimation in the Gaussian graphical model. Biometrika, 94(1), 19-35.

[66] Sun, L., Patel, R., Liu, J., Chen, K., Wu, T., Li, J., Reiman, E., Ye, J., 2009. Mining brain region connectivity for Alzheimer's disease study via sparse inverse covariance estimation. Proceedings of Knowledge Discovery and Data Mining Conference (KDD), 1335-1344.

[67] Li, H., Gui, J., 2006. Gradient directed regularization for sparse Gaussian concentration graphs, with applications to inference of genetic networks. Biostatistics, 7(2), 302–317.

[68] Schafer, J., Strimmer, K., 2005. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Statistical Applications in Genetics and Molecular Biology, 4(1).

[69] Bickel, P. J., Levina, E., 2008. Regularized estimation of large covariance matrices. Annals of Statistics, 36(1), 199–227.

[70] Levina, E., Rothman, A. J., Zhu, J., 2008. Sparse estimation of large covariance matrices via a nested Lasso penalty. Annals of Applied Statistics, 2(1), 245–263.

[71] Meinshausen, N., Bühlmann, P., 2006. High dimensional graphs and variable selection with the Lasso. Annals of Statistics, 34(3), 1436–1462.

[72] Peng, J., Wang, P., Zhou, N., Zhu, J. Partial correlation estimation by joint sparse regression models, 2009. Journal of the American Statistical Association, 104, 735-746.

[73] Friedman, J. H., Hastie, T. and Tibshirani, R., 2010. Applications of the lasso and grouped lasso to the estimation of sparse graphical models.

[74] Hsieh, C.-J., Sustik, M. A., Dhillon, I. S., Ravikumar, P. K., 2011. Sparse Inverse Covariance Matrix Estimation Using Quadratic Approximation. Advances in Neural Information Processing Systems (NIPS).

[75] Hsieh, C.-J., Sustik, M. A., Dhillon, I. S., Ravikumar, P. K., Poldrack, R. A., 2013. BIG & QUIC: Sparse Inverse Covariance Estimation for a Million Variables. Advances in Neural Information Processing Systems (NIPS).

[76] Hanauer, D.A., Rhodes, D.R., Chinnaiyan, A.M., 2009. Exploring clinical associations using "-omics" based enrichment analyses. Plos One.

[77] Kolar, M., Song, L., Ahmed, A., Xing, E.P., 2010. Estimating time-varying networks. Annals of Applied Statistics, 4(1), 94-123.

[78] Lafferty, J., Liu, H, Wasserman, L., 2012. Sparse nonparametric graphical models. Statistical Science, 27(4), 519-537.

[79] Liu, H., Han, F., Yuan, M., Lafferty, J., Wasserman, L., 2012. High-dimensional semiparametric Gaussian copula graphical models. Annals of Statistics, 40(4), 2293-2326.

[80] Rokach, L., Maimon, O., 2008. Data Mining with Decision Trees: Theory and Applications. World Scientific Publishing.

[81] Tibshirani, R., 1996. Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society Series B, 58(1), 267–288

[82] Hastie, T.; Tibshirani, R. and Friedman, J., 2008, Elements of Statistical Learning, 2nd Edition, Springer.

[83] Liu, J., Chen, J.H., Ye, J., 2009. Large-scale sparse logistic regression. The fifteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 547-556.

[84] Therneau, T.M., Atkinson, E.J., 2013. An introduction to recursive partitioning using the RPART routines, technical report.

[85] Mansour, Y., 2000. Generalization bounds for decision tree. COLT.

[86] Golea, M., Bartlett, P.L., Lee, W.S., Mason, L., 1997. Generalization in decision tree and DNF: does size matter? NIPS.

[87] Pichuka, C., Bapi, R.S., Bhagvati, C., Pujari, A.K., Deekshatulu, B.L., 2007. A Tighter Error Bound for Decision Tree Learning Using PAC Learnability, IJCAI.

[88] Murthy, S.K., 1997. Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey. Data Mining and Knowledge Discovery.

[89] Ehrenfeucht, A., Haussler, D., 1989. Learning decision trees from random examples. Information and Computation, 82(2), 231-246.

[90] Fayyad, U.M., Irani, K.B., 1990. What should be minimized in a decision tree? AAAI.

[91] Su, J., Zhang, H., 2006. A fast decision tree learning algorithm. AAAI.

[92] Friedman, J.H., Popescu, B.E., 2008. Predictive learning via rule ensembles. Ann. Appl. Stat., 2(3), 916-954.

[93] Glass, D., Lisk, C., Stensland, J., 2012. Refining the hospital readmissions reduction program. Washington, DC: Medicare Payment Advisory Commission.

[94] Pfohl, M., Koch, M., Enderle, M.D., Kühn, R., Füllhase, J., Karsch, K.R., Häring. H.U., 1999. Paraoxonase 192 Gln/Arg gene polymorphism, coronary artery disease, and myocardial infarction in type 2 diabetes. Diabetes, 48(3), 623-627.

[95] Shanmuga, S.P., Padma, S., Kumar, H., Nair, V., Kumar, S., 2007. Role of 99mTc MDP bone and 67Gallium imaging in evaluation of diabetic osteopathy. The Foot, 17(2), 94-101.

[96] Berlin, I., 2008. Smoking-induced metabolic disorders: a review. Diabetes & Metabolism, 34(4), 307-314.

[97] Danaher P, Wang P, Witten, D., 2013. The joint graphical lasso for inverse covariance estimation across multiple classes. To appear in Journal of the Royal Statistical Society, Series B.

[98] Liu, J., Zhao, Z., Wang, J., Ye, J., 2014.Safe Screening with Vibrational Inequalities and Its Applications to Lasso, ICML.

[99] Wang, J., Zhou, J., Wonka, P., Ye, J., 2013.Lasso Screening Rules via Dual Polytope Projection, NIPS.

[100] Achieving operating room efficiency through process integration, 2003. Healthc Financ Manage 57(3), S1-S7.

[101] Barnoon, S., Wolfe, H., 1968. Scheduling a multiple operating room system: A simulation approach. Health Serv Res., 3(4), 272-285.

[102] Robinson, G., H., Wing, P., Davis, L. E., 1968. Computer simulation of hospital patient scheduling systems. Health Serv Res., 3(2), 130-141.

[103] Fei, H., Meskens, N., Chu., C., 2010. A planning and scheduling problem for an operating theatre using an open scheduling strategy. Computers & Industrial Engineering, 58(2), 221-230.

[104] Huang, Y. L., Hancock, W. M., Herrin, G. D., 2012. An alternative outpatient scheduling system: Improving the outpatient experience. IIE Transactions on Healthcare Systems Engineering, 2(2), 97-111.

[105] Huele, C. V., Vanhoucke, M., 2014. Analysis of the Integration of the Physician Rostering Problem and the Surgery Scheduling Problem. Journal of Medical Systems, 38(6), 1-16.

[106] Meskens, N., Duvivier, D., Hanset, A., 2013. Multi-objective operating room scheduling considering desiderata of the surgical team. Decision Support Systems, 55(2), 650-659.

[107] Lamiri, M., Xie, X., Dolgui, A., Grimaud, F., 2008. A stochastic model for operating room planning with elective and emergency demand for surgery. European Journal of Operational Research, 185(3), 1026-1037.

[108] Addis, B., Carello, G., Tanfani, E., 2014. A robust optimization approach for the Advanced Scheduling Problem with uncertain surgery duration in Operating Room Planning-an extended analysis. Working paper.

[109] Denton, B. T., Miller, A. J., Balasubramanian, H. J., Huschka, T. R., 2010. Optimal Allocation of Surgery Blocks to Operating Rooms Under Uncertainty. Operations Research, 58(4), 802-816.

[110] May, J. H., Spangler, W. E., Strum, D. P., Vargas, L. G., 2011. The Surgical Scheduling Problem: Current Research and Future Opportunities. Production and Operations Management, 20(3), 392-405.

[111] Delignette-Muller, M. L., Dutang, C., Pouillot, R., Denis, J. B., 2014. Package 'fitdistrplus'.

[112] Etzioni, D. A., Liu, J. H., Maggard, M. A., Ko, C. Y., 2003. The aging population and its impact on the surgery workforce. Ann Surg., 238(2), 170-177.

[113] Lindholm, M, Hargraves, J. L., Ferguson, W. J., Reed, G., 2012. Professional Language Interpretation and Inpatient Length of Stay and Readmission Rates. Journal of General Internal Medicine, 27(10), 1294-1299.

APPENDIX A: COPYRIGHT PERMISSIONS

Below is permission for the use of material in Chapter 3.

Copyright Clearance Center RightsLink®

Home | Account Info | Help

Title: Learning high-dimensional networks with nonlinear interactions by a novel tree-embedded graphical model

Publication: Pattern Recognition Letters

Publisher: Elsevier

Date: 1 November 2014

Copyright © 2014 Elsevier B.V. All rights reserved.

Logged in as:
Yazhuo Liu
Account #: 3000916793

LOGOUT

Order Completed

Thank you very much for your order.

This is a License Agreement between Yazhuo Liu ("You") and Elsevier ("Elsevier"). The license consists of your order details, the terms and conditions provided by Elsevier, and the payment terms and conditions.

Get the printable license.

License Number	3635010571263
License date	May 23, 2015
Licensed content publisher	Elsevier
Licensed content publication	Pattern Recognition Letters
Licensed content title	Learning high-dimensional networks with nonlinear interactions by a novel tree-embedded graphical model
Licensed content author	None
Licensed content date	1 November 2014
Licensed content volume number	49
Licensed content issue number	n/a
Number of pages	7
Type of Use	reuse in a thesis/dissertation
Portion	full article
Format	electronic
Are you the author of this Elsevier article?	Yes
Will you be translating?	No
Title of your thesis/dissertation	Patient populations, clinical associations, and system efficiency in healthcare delivery system
Expected completion date	Jul 2015
Estimated size (number of pages)	90
Elsevier VAT number	GB 494 6272 12
Permissions price	0.00 USD
VAT/Local Sales Tax	0.00 USD / 0.00 GBP
Total	0.00 USD

ORDER MORE... | CLOSE WINDOW

Copyright © 2015 Copyright Clearance Center, Inc. All Rights Reserved. Privacy statement. Terms and Conditions.
Comments? We would like to hear from you. E-mail us at customercare@copyright.com

Below is permission for the use of Table 5, Table 6, and Figure 2.

.

Copyright Clearance Center RightsLink®

Home | Account Info | Help

Title: Bounceback branchpoints: Using conditional inference trees to analyze readmissions

Publication: Surgery

Publisher: Elsevier

Date: October 2014

Copyright © 2014 Elsevier Inc. All rights reserved.

Logged in as:
Yazhuo Liu
Account #:
3000916793

LOGOUT

Order Completed

Thank you very much for your order.

This is a License Agreement between Yazhuo Liu ("You") and Elsevier ("Elsevier"). The license consists of your order details, the terms and conditions provided by Elsevier, and the payment terms and conditions.

Get the printable license.

License Number	3635010924855
License date	May 23, 2015
Licensed content publisher	Elsevier
Licensed content publication	Surgery
Licensed content title	Bounceback branchpoints: Using conditional inference trees to analyze readmissions
Licensed content author	None
Licensed content date	October 2014
Licensed content volume number	156
Licensed content issue number	4
Number of pages	7
Type of Use	reuse in a thesis/dissertation
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	3
Format	electronic
Are you the author of this Elsevier article?	Yes
Will you be translating?	No
Original figure numbers	Table I, Table II, and Fig
Title of your thesis/dissertation	Patient populations, clinical associations, and system efficiency in healthcare delivery system
Expected completion date	Jul 2015
Estimated size (number of pages)	90
Elsevier VAT number	GB 494 6272 12
Permissions price	0.00 USD
VAT/Local Sales Tax	0.00 USD / 0.00 GBP
Total	0.00 USD

ORDER MORE... | CLOSE WINDOW

Copyright © 2015 Copyright Clearance Center, Inc. All Rights Reserved. Privacy statement. Terms and Conditions.
Comments? We would like to hear from you. E-mail us at customercare@copyright.com

Below is permission for the use of Table 1, Table 2, Table 3 and Table 4.

Wolters Kluwer Rights and Permissions <journalpermissions@lww.com> Fri, May 8, 2015 at 5:26 AM
Reply-To: Wolters Kluwer Rights and Permissions <journalpermissions@lww.com>
To: yazhuoliu@mail.usf.edu

Recently you requested personal assistance from our on-line support center. Below is a summary of your request and our response.

If you are receiving this in response to a request you made, a summary is below. If you have not made a request, the following is a communication on behalf of your LWW Sales Representative.

Thank you for allowing us to be of service to you.

Subject

[Question] Permission to Reuse Tables/Figures in My PhD Dissertation.

Discussion Thread

Response Via Email (Delayna S.) 05/08/2015 05:26 AM

Dear Yazhuo Liu,

Thank you for your email.

As you are the second author of this article, you are in fact entitled to reuse it within your dissertation without formally requesting permission (please see the attached author's permissions document for further information).

We just ask that if you wish to reuse the entire article, you do not make any modifications to it.

Kind regards,
Delayna

Delayna Spencer
Permissions Assistant

Wolters Kluwer
250 Waterloo Road, London, SE1 8RD, England, United Kingdom

+44 (0)207 981 0518 Direct Line
+44 (0)207 981 0562 Fax
Delayna.Spencer@wolterskluwer.com
www.wolterskluwerhealth.com

Confidentiality Notice: This email and its attachments (if any) contain confidential information of the sender. The information is intended only for the use by the direct addressees of the original sender of this email. If you are not an intended recipient of the original sender (or responsible for delivering the message to such person), you are hereby notified that any review, disclosure, copying, distribution or the taking of any action in reliance of the contents of and attachments to this email is strictly prohibited. If you have received this email in error, please immediately notify the sender at the address shown herein and permanently delete any copies of this email (digital or paper) in your possession.