A STUDY OF UNPLANNED 30-DAY HOSPITAL READMISSIONS IN THE UNITED STATES: EARLY PREDICTION AND POTENTIALLY MODIFIABLE RISK FACTOR IDENTIFICATION

A Dissertation

presented to

the Faculty of the Graduate School

at the University of Missouri-Columbia

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

---

by

PENG ZHAO

Dr. Illhoi Yoo, Dissertation Supervisor

MAY 2020

The undersigned, appointed by the dean of the Graduate School, have examined the dissertation entitled

A STUDY OF UNPLANNED 30-DAY HOSPITAL READMISSIONS IN THE UNITED STATES: EARLY PREDICTION AND POTENTIALLY MODIFIABLE RISK FACTOR IDENTIFICATION

presented by Peng Zhao,

a candidate for the degree of Doctor of Philosophy and hereby certify that, in their opinion, it is worthy of acceptance.

Dr. Illhoi Yoo

Dr. Sue Boren

Dr. Mihail Popescu

Dr. Abu Mosa

Dr. Lincoln Sheets

DEDICATION

This dissertation is dedicated to my mother (Jingrui Zhang, 张景瑞), my father (Ziying Zhao, 赵子英), my sister (Hongyan Zhao, 赵鸿雁), my brother in law (Bo Jin, 金波), and my niece (Yihan Jin, 金逸涵), for their endless love, support and encouragement.

ACKNOWLEDGMENTS

First of all, I would like to thank my advisor, Dr. Illhoi Yoo for his patient and careful guidance of my Ph.D. studies. This project would not have been possible without his inspiration and support.

Secondly, I want to thank my committee members, Dr. Sue Boren, Dr. Mihail Popescu, Dr. Abu Mosa, and Dr. Lincoln Sheets for their continued help in my research.

I am also thankful to Dr. Chi-Ren Shyu, Dr. Gavin Conant, Dr. Timothy Matisziw, Mr. Robert Sanders, and Ms. Tracy Pickens for their advice and help.

Finally, thanks to all my fellow friends at the MU Institute for Data Science and Informatics who made me not lonely on this long journey.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

LIST OF ABBREVIATIONS

ADTree	Alternating Decision Tree
AHRQ	Agency for Healthcare Research and Quality
AIC	Akaike Information Criterion
AMI	Acute Myocardial Infarction
AUC	Area under the Receiver Operating Characteristic Curve
B	"Need to Improve" Satisfaction
BMI	Body Mass Index
CABG	Coronary Artery Bypass Graft
CCS	Clinical Classification Software
CI	Confidence Interval
CMS	Centers for Medicare and Medicaid Services
COPD	Chronic Obstructive Pulmonary Disease
CPT	Current Procedural Terminology
DRG	Diagnosis-Related Group
DX	Diagnosis
EHR	Electronic Health Record
G	"No Need to Improve" Satisfaction
HCPCS	Healthcare Common Procedure Coding System
HCUP	Healthcare Cost and Utilization Project
HF	Heart Failure
HL	Hosmer-Lemeshow
HRRP	Hospital Readmission Reduction Program
ICD-9-CM	International Classification of Diseases, Ninth Revision, Clinical Modification
ICD-10-CM	International Classification of Diseases, Tenth Revision, Clinical Modification
ICD-10-PCS	International Classification of Diseases, Tenth Revision, Procedure Coding System
IPPS	Inpatient Prospective Payment System

MOI	Missouri Orthopedic Institute
N	No
NHS	National Health Service
NRD	Nationwide Readmissions Database
OR	Odds Ratio
PN	Pneumonia
ROC	Receiver Operating Characteristic
THA/TKA	Total Hip or Knee Arthroplasty
XGBoost	Extreme Gradient Boosting
YD	Yes, Definitely
YS	Yes, Somewhat

ABSTRACT

Unplanned hospital readmissions greatly impair patients' quality of life and have imposed a significant economic burden on American society. The pressure to reduce costs and improve healthcare quality has triggered the development of readmission reduction interventions. However, existing solutions focus on complementing inpatient care with enhanced care transition and post-discharge interventions, which are initiated near or after discharge when clinicians' impact on inpatient care is ending. Preventive intervention during hospitalization is an under-explored area, which holds the potential for reducing readmission risk. Nevertheless, it is challenging for clinicians to predict readmission risk at the early stage of inpatient care because little data is available. Existing readmission predictive models tend to incorporate variables whose values are only available near or after discharge. As a result, these models cannot be used for the early prediction of readmission. Another challenge is that there is no universal solution to reduce readmissions during hospitalization. Patients can be readmitted for any reason, and their heterogeneous social and clinical factors can further complicate the planning of interventions. The objective of this project was to improve the timeliness of readmission preventive intervention through a data-driven approach. A systematic review of the literature was performed to collect reported risk factors for unplanned 30-day hospital readmission. Using various predictive modeling and exploratory analysis methods, we have developed an early prediction model of readmission and have identified potentially modifiable readmission risk factors, which may be used to guide the development of readmission preventive interventions during hospitalization for different patients.

Chapter 1 Introduction

1.1 Unplanned 30-Day Hospital Readmissions

An unplanned hospital readmission means that a patient unexpectedly returns to the hospital after being discharged from a previous admission (index admission) within a specific time interval (e.g., 30 days). It is an undesired healthcare outcome that impairs patients' quality of life due to prolonged illness and emotional stress [1]. The substantial costs associated with unplanned hospital readmissions have imposed a significant economic burden on American society. During July 2015 to June 2016, 15.2% of Medicare beneficiaries experienced unplanned 30-day hospital readmissions [2]. Every year, unplanned hospital readmissions are estimated to account for $17.4 billion in Medicare expenditure [3].

In this project, we adopted the Centers for Medicare and Medicaid Services' (CMS) definition of a hospital readmission as "an admission to an acute care hospital within 30 days of a discharge from the same or another acute care hospital" [4]. The CMS considers all unplanned readmissions (all-cause) in the readmission measure [4]. The reason is that unplanned readmissions, regardless of the cause, are adverse events and they should be considered when measuring quality [4]. We used the planned readmission identification criteria [5] developed by the CMS to identify unplanned 30-day hospital readmissions (hereafter, referred to as "readmissions" unless otherwise stated).

1.2 The Hospital Readmission Reduction Program

Reducing readmission has captivated policymakers as a goal that improves healthcare quality and reduce costs [6]. Since 2009, the CMS has been publicly reporting the hospital-level risk-standardized readmission rates on the Hospital Compare [7] website, which allows patients to compare hospitals with government ratings [8]. In 2012, the Affordable Care Act [9] implemented the Hospital Readmission Reduction Program (HRRP) [4]. Under this program, hospitals participating in the inpatient prospective payment system (IPPS) will be assessed using their 30-day hospital readmission rates following several eligible conditions and surgeries in initial hospitalizations, adjusted for age, sex, and comorbidities [4]. The readmission rates will be compared with averages of national readmission rates, and hospitals with excessive readmission rates will receive percentage reduction of total Medicare payments [4]. As of the fiscal year 2020, eligible conditions and surgeries include acute myocardial infarction (AMI), chronic obstructive pulmonary disease (COPD), heart failure (HF), pneumonia (PN), coronary artery bypass graft (CABG), and total hip or knee arthroplasty (THA/TKA) [4]. They were selected due to high cost, substantial morbidity and mortality, and marked performance variations across hospitals [10]. Table 1.1 shows the information of the HRRP penalties in fiscal years 2013 to 2020 [11–14]. The percentage of penalized hospitals and the amount of penalty have been increasing since the implementation of the HRRP.

Table 1.1 The HRRP penalties in fiscal years 2013 to 2020.

Fiscal Year	2013	2014	2015	2016	2017	2018	2019	2020
Eligible conditions & surgeries	AMI HF PN	AMI HF PN	AMI HF PN COPD THA TKA	AMI HF PN COPD THA TKA	AMI HF PN COPD THA TKA CABG	AMI HF PN COPD THA TKA CABG	AMI HF PN COPD THA TKA CABG	AMI HF PN COPD THA TKA CABG
Percent of penalized hospitals	64%	66%	78%	78%	79%	79%	82%	83%
Estimated penalty	$290 M	$227 M	$428 M	$420 M	$528 M	$564 M	$566 M	$563 M

1.3 Interventions to Reduce Readmissions

The pressure to reduce costs and improve healthcare quality has triggered the development of readmission reduction interventions, which can be classified into pre-discharge interventions, transition interventions, and post-discharge interventions based on the timing of intervention. Hansen et al. [6] conducted a systematic review of 43 readmission preventive intervention studies and identified 12 categories of interventions (Table 1.2).

Table 1.2 Categories of interventions to reduce readmissions.

Pre-discharge Interventions	Transition Interventions	Post-discharge Interventions
• Patient education • Discharge planning • Medication reconciliation • Appointment scheduled before discharge	• Transition coach • Patient-centered discharge instructions • Provider continuity	• Timely follow-up • Timely primary care provider communication • Follow-up telephone call • Patient hotline • Home visit

1.4 Problem Statement

Existing readmission reduction interventions, especially transition interventions and post-discharge interventions, focus on passively complementing inpatient care with enhanced services, whose planning, implementation, and monitoring can be resource-intensive [15]. In addition, no single or bundle of these interventions were found to be reliable in readmission reduction according to the review by Hansen et al. [6]. Another disadvantage is that these interventions can hardly impact the quality improvement of inpatient care because they are mostly initiated near or after discharge when clinicians' impact on inpatient care is ending.

Preventive intervention during hospitalization is an under-explored area, which holds the potential for reducing readmission risk. It has been shown that unplanned hospital readmissions are related to inadequate or substandard inpatient care, such as undertreatment, premature discharge, healthcare-associated complications, and medical errors [16–20]. Early interventions, such as early discharge planning, are effective in reducing readmissions [21]. However, it is impractical to deliver readmission preventive interventions to all patients due to restricted healthcare resources. Predictive modeling is an efficient method to optimize the allocation of valuable clinical resources by stratifying patients' readmission risk and target the delivery of preventive interventions to patients at high risk [22]. Evidence has shown that applying interventions to high-risk patients can reduce 30-day hospital readmission risk by 11-28% [23–25].

Nevertheless, the majority of reported hospital readmission predictive models have limited clinical values because they require variables whose values only become

completely available at discharge [26]. For example, the LACE index [27] and the HOSPITAL score [28] are the most widely used readmission prediction models in US healthcare settings. Both of them require the length of inpatient stay. The HOSPITAL score also requires two lab test results at discharge. These variables can only be available near discharge when clinicians can no longer provide impactful care. It is essential to perform early risk assessments of high-risk patients so that clinicians can deliver timely preventive interventions at the early stage of hospitalization [29].

Another challenge is that there are few interventions available during hospitalization. It can be seen from Table 1.2 that most of the existing interventions are designed to enhance care transitions and follow-ups. The basic assumption is that poor care coordination and poor follow-up care after discharge are the major causes of hospital readmissions [30–32]. Nevertheless, according to a survey of patients who experienced 30-day unplanned hospital readmissions about their experience of discharge and post-discharge care, more than 74% of the 530 eligible respondents reported that they were readmitted even though they had a good knowledge of the discharge plan and the post-discharge care, including self-care, medications, and communications with doctors [30]. One possible reason for this discrepancy is that readmissions are not only impacted by the care transition and the care after discharge, but also the whole episode of inpatient care [33].

1.5 Overall Objective and Aims

The overall objective of this project was to improve the timeliness of readmission preventive intervention by enabling early prediction of readmission risk and

recommending potentially modifiable risk factors associated with readmission. We have achieved this objective by accomplishing the following three specific aims.

1.5.1 Aim 1

The first aim was to investigate the risk factors for unplanned 30-day hospital readmissions. We performed a systematic review of 13 eligible studies and identified 34 highly generalizable risk factors. Chapter 2 presents the results. These risk factors were used to guide the selection of variables of the early predictive model in Aim 2.

1.5.2 Aim 2

The second aim was to build an early predictive model of readmission to identify high-risk patients at the early stage of hospitalization. We created features from patients' medical history data within one year before hospitalization and index admission's data that can be available in the electronic health record (EHR) within 24 hours. We applied various statistical and machine learning algorithms for readmission risk predictive modeling and developed a model with the performance better than reported models. In addition, we identified 14 novel risk factors and two novel protective factors of readmission by multivariate analysis. The results are shown in Chapter 3.

1.5.3 Aim 3

The third aim was to identify potentially modifiable risk factors associated with readmission. These risk factors can potentially be used as the target to plan and deliver interventions. To study the association between the potential change of a risk factor (e.g., a medical condition) and the change of readmission status, we compared the same patients'

different index admissions, and our approach consists of a hybrid of data mining and statistical methods. We identified association rules associated with the change of readmission status and showed the results in Chapter 4. Because each association rule represents a patient subgroup, clinicians can use it to customize interventions for patients falling in the subgroup. Furthermore, we applied the same method to study factors associated with the change of orthopedic patient satisfaction and identified a novel patient-provider sex concordance pattern that can be potentially used to improve orthopedic patient satisfaction. The results are displayed in Chapter 5.

1.6 Significance

Unplanned hospital readmissions have attracted a lot of attention due to the negative influence on patients' quality of life and substantial costs. The penalties from the HRRP have intensified efforts from the entire healthcare industry to reduce unplanned hospital readmissions. As described in section 1.4 of this chapter, existing readmission preventive solutions are limited in the timeliness of readmission risk predictions and interventions. If clinicians can predict patients' readmission risk at the early stage of hospitalization and know their modifiable risk factors, they can potentially better plan and deliver enhanced treatment plans. This project can potentially help to solve this problem by creating an early prediction model of readmission and identifying potentially modifiable risk factors associated with readmission. Furthermore, this project has potentially positive financial impacts by reducing costs caused by readmissions themselves and the HRRP penalties. According to the report by the Medicare Payment Advisory Commission, 12% of readmissions are potentially preventable, and the CMS could save $1 billion every year for reducing 10% of these readmissions [32].

1.7 Innovation

This project has three innovations. First, it represents a shift in the timing of readmission preventive interventions. Most readmission reduction programs focus on transition care and post-discharge care, whereas our work can potentially enable clinicians to identify high-risk patients and plan interventions for them at the early stage of inpatient care. Second, we have identified 14 novel risk factors and two novel protective factors of readmission. To our knowledge, they have never been reported. They would facilitate clinical research to further understand the causes of readmission. Third, we developed a novel data analysis method based on association rule mining and statistical methods to analyze the differences between the same patients' different inpatient visits. This method allowed us to identify potentially modifiable risk factors associated with the change of readmission status. With the same method, we found a novel patient-provider sex concordance pattern associated with the change of orthopedic patient satisfaction. This pattern can be potentially used to improve orthopedic patient satisfaction.

1.8 Outcomes

This project has produced four journal articles. Chapter 2 has been published in the Journal of Health and Informatics in 2017 [34]. Chapters 3 to 5 have been submitted as three journal articles and are under review at the time of preparing this dissertation.

Chapter 2 A Systematic Review of Risk Factors for Unplanned 30-day Hospital Readmission

We have published this systematic review in the Journal of Health and Informatics in 2017 [34]. This chapter adopts its main content with minor modifications to reflect our latest findings.

2.1 Background

Recent years have seen a growing body of literature on hospital readmissions with the goal of improving healthcare quality and lowering cost. Predictive modeling of readmissions is one of the most common study types to help providers better identify high-risk patients. Unfortunately, studies in this area are highly fragmented, especially in target populations. The study outcomes span from models that are specific to populations with particular diseases or surgeries to general purpose models applicable to all patients. As of the fiscal year 2020, the HRRP in the United States only considers the index conditions or surgeries of acute myocardial infarction, heart failure, pneumonia, chronic obstructive pulmonary disease, elective total hip or knee arthroplasty, and coronary artery bypass graft in the calculation of readmission penalties due to their high prevalence and cost [4,35]. Largely spurred by the HRRP, many studies have focused on readmissions occurring after the index admissions for these conditions or surgeries only. The choice between condition-specific and all-condition readmission models has long been under debate. However, condition-specific models have been criticized for the poor generalizability, especially in

patients with multiple conditions [36,37]. In addition, it has been reported that 58.5% of unplanned readmissions were clinically irrelevant to the index admissions [38].

Attempts to predict readmissions were further complicated by the lack of consensus on data inclusion criteria. For instance, most studies focused on unplanned readmissions while some others included all available readmissions without removing scheduled readmissions. The definitions of unplanned readmissions were also highly inconsistent. Some studies restricted unplanned readmissions to occur in certain hospital departments or specialties [39,40] and some identified them by diagnosis-related groups (DRG) [41,42]. Unplanned readmissions were further classified as either potentially avoidable or unavoidable. Readmissions due to progressions of existing conditions or newly developed conditions after discharge are deemed unavoidable [43,44]. It has been argued that including unavoidable readmissions in quality measures is unfair because they are not directly related to the quality of healthcare services during index admissions [43]. However, there is no agreement at present on the criteria to identify avoidable readmissions. In many studies, the avoidable readmissions were determined by medical experts and the inclusion eligibility can be subjective [45]. According to a systematic review of 34 articles in 2011, the measured proportions of avoidable readmissions varied from 5% to 79% [43].

To exacerbate the situation, diverse time frames were used to capture readmissions. In a systematic review of 26 readmission prediction models developed in six countries, the intervals between discharges and readmissions ranged from 14 days to four years [26]. The CMS in the United States adopted the 30-day time window [4]. In the United Kingdom, both 28-day [46] and 30-day [47] periods were used by the National Health Service (NHS) to measure readmission rates. 30-day is currently the most used time frame globally. The

possible reasons are that older patients are more vulnerable during this period [48] and readmissions occurring within 30 days are more likely influenced by the quality of care [33].

Given the complex nature of hospital readmissions, it is challenging to conduct meaningful readmission prediction studies without good knowledge of existing evidence from both domestic and global research communities. However, due to the heterogeneous target populations and inconsistent definitions of readmissions, the outcomes of some studies can be hardly generalized to other studies [49]. The purpose of this study was to identify the generalizable study outcomes of readmission predictions from the risk factor level to guide the selection of baseline predictor variables in different readmission studies. Especially, we are interested in the risk factors for unplanned 30-day all-cause hospital readmissions due to the better-validated time frame and the broader target.

In the past few years, several attempts have been made to review risk factors or predictor variables for hospital readmissions, yet none of them have focused on generalizability. The review by Vest et al. [50] in 2010 was limited to US studies from 2000 to 2009 only and the time frame varied from seven-day to six-month for all-cause readmissions. The focus of the review by Kansagara et al. [26] in 2011 was on readmission prediction models derived in developed countries before 2011. Predictor variables of the reviewed models were tabulated without differentiating condition-specific and all-cause models. Zhou et al. [22] also placed emphasis on reviewing readmission prediction models developed between 2011 and 2015. The significant predictor variables were summarized without further analysis. Besides, many studies have reviewed readmission risk factors for specific conditions or surgeries. To the best of our knowledge, this is the first systematic

review of highly generalizable risk factors for unplanned 30-day all-cause hospital readmissions.

2.2 Materials and Methods

2.2.1 Data Source and Search Strategy

A literature search was performed in PubMed to identify articles relating to the risk factors for unplanned 30-day all-cause hospital readmissions. The search keywords have four components reflecting the interest of this review: "unplanned", "30-day", "hospital readmission", "risk factors". "All-cause" was not included in the keyword because some all-cause readmission studies do not explicitly mention their scopes. Synonyms and hyphenations were included to account for variations in different studies. Wildcards were used to match the verb and noun forms of "readmission". The logical relationships among the search keywords are shown in Figure 2.1.

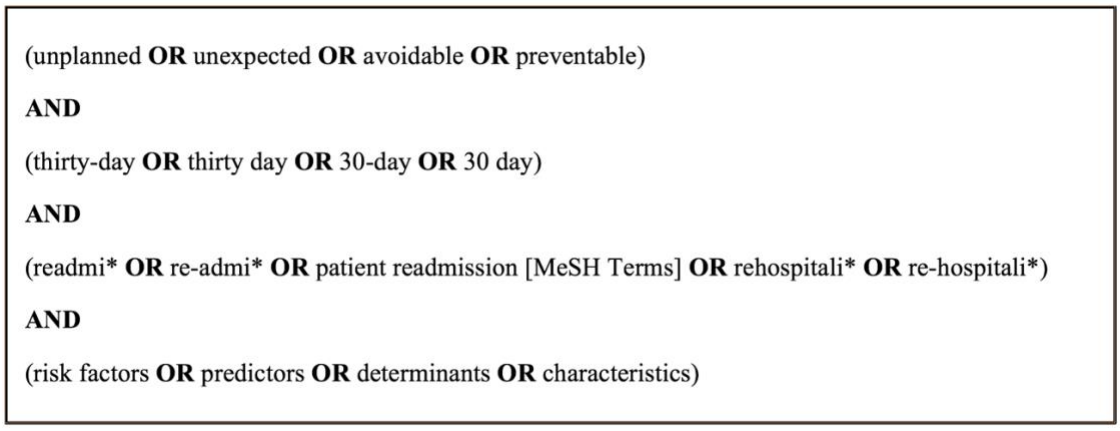


Figure 2.1 Logical relationships among the search keywords.

2.2.2 Study Inclusion and Exclusion Criteria

In this study, only peer-reviewed articles written in English were considered. We included articles focusing on identifying statistically significant predictor variables or risk factors for 30-day unplanned all-cause hospital readmissions. Articles were excluded if they met any of the following criteria: (1) The readmission time frame is other than 30 days, such as 90-day readmission, (2) Studies focusing on planned readmissions or not differentiating planned and unplanned readmissions, (3) Studies specific to narrow patient populations with particular medical conditions or underwent certain surgeries, (4) The study outcome is more than 30-day unplanned all-cause hospital readmission, such as mortality in combination with 30-day unplanned all-cause hospital readmission, (5) Studies of pediatric and newborn readmissions. Pediatric and newborn readmissions were filtered out because the risk factors may be distinct from adult readmissions [26,50] and the readmissions could be influenced by parental factors [50,51]. To reduce redundancy and bias, external validations of existing prediction tools were removed and only the original articles of the cited tools were included if eligible.

2.2.3 Data Extraction Process

The characteristics of the studies, including the publication year, study region, data source, study design, cohort definition, definition of unplanned readmissions, analysis method, predictor variables, and risk factors ($P < 0.05$) were extracted from all the included studies. The risk factors were summarized by category and were grouped if they share the same corresponding predictor variable. The number of studies that analyzed predictor variables and the number of studies that found them significant were recorded.

2.2.4 Generalizability Assessment

In this study, we define "generalizability" as the capability of being applied to other hospital readmission prediction studies regardless of target populations and residing places. Although the studies specific to narrow populations have been filtered out during the article selection step, some identified risk factors may be still tied to a sub-population. For example, studies with Medicare patients are less generalizable because those patients are 65 years old or older in the United States. Also, some risk factors identified in one place may not work in other places if they are closely related to the unique local healthcare systems (e.g., insurance, medical social welfare). In addition, it may be impractical to apply some risk factors to other types of studies due to the difference in study exposures related to designs. As a result, we chose to assess the generalizability of risk factors by three questions: (1) Whether a risk factor is specific to a narrow population or not, (2) Whether it is specific to a place or not, (3) Whether it is specific to a study exposure related to one particular study design or not. If a risk factor is not specific to any of them, we deem that the risk factor is generalizable.

2.3 Results

2.3.1 Study Selection

Figure 2.2 shows the process of identifying eligible articles. The initial query was performed on July 21, 2017 and returned 370 articles. After removal of one duplicated article and two non-English articles, the remaining 367 articles were reviewed based on titles and abstracts. 331 of them met the exclusion criteria and were filtered out. The remaining 36 articles were then reviewed in full text. One article was removed because it

is not a peer-reviewed article. Four studies were excluded from the list because they are external validations of two existing readmission prediction models without major modifications (two articles validated the HOSPITAL score [28], one article validated the LACE index [27], and one article validated both the HOSPITAL score and LACE index). The original article of HOSPITAL score was included in this review while the LACE index article was not because the study outcome was both mortality and 30-day unplanned readmission. Four articles were specific to certain diagnoses and thus were removed. Five studies were filtered out because they did not differentiate unplanned readmissions. Nine articles were excluded because they did not report statistically significant predictor variables or risk factors. The remaining 13 highly relevant articles were reviewed.

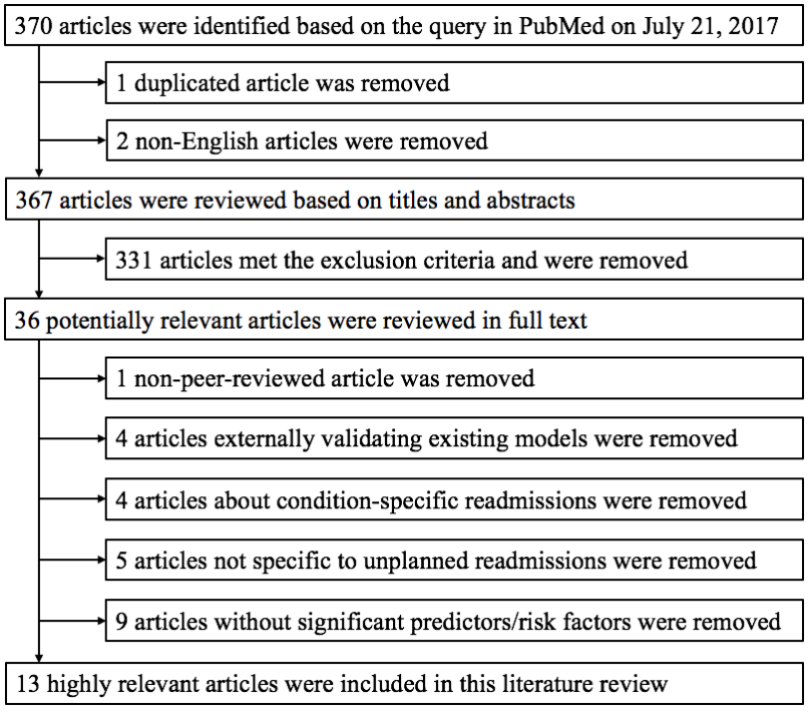


Figure 2.2 Trial flow diagram of the process to identify eligible articles.

2.3.2 Data Extraction

The literature on this topic is very recent. Although we did not intentionally limit the publication date, the earliest eligible article was published in 2009, reflecting a growing interest in predicting 30-day unplanned all-cause readmissions in the past decade. Of the 13 studies, over half (7/13) are based in the United States, two in Israel, two in Singapore, one in Sweden, and one in Taiwan. The majority (12/13) of the studies are retrospective and only one study [52] adopted the prospective design. Multivariate logistic regression is the most used analysis method (12/13) to identify significant predictor variables and only one study [53] used Poisson regression.

Studies are highly heterogeneous in data type and data sources. Two studies [54,55] used claims data and five studies [28,40,42,56,57] used clinical and/or administrative data from EHR. Four studies [39,41,52,58] combined data from various sources, including proprietary EHR, validated questionnaires, hospital information systems, Veterans Affairs database, and Medicare dataset. One article [59] studied state-level discharge summary data and one study [53] retrospectively analyzed the control group of a clinical trial.

The definitions of unplanned readmissions are also very distinct. Four studies [52,56,57,59] directly used the data of unplanned readmissions without any definitions. Two studies [39,40] only included readmissions to emergency departments within 30 days of discharge because emergency department visits are not scheduled in advance. Three studies [41,42,54] excluded planned readmissions based on Clinical Classification Software (CCS) [60] or Diagnosis-Related Group (DRG) [61] codes, including transplantations, psychiatric issues, maintenance chemotherapy, dental procedures,

pregnancy-related procedures, and other planned procedures. Two studies [53,58] excluded admissions to the specialties of obstetrics, gynecology, dentistry, otolaryngology, ophthalmology, orthopedic surgery, general surgery, or psychiatry. One study [55] excluded admissions with a principal diagnosis of cancer because cancer patients may have planned stays for cancer treatments. One study [28] separated readmissions into potentially avoidable and unavoidable based on administrative data with a validated algorithm SQLape [62]. Unavoidable readmissions include planned readmissions and any unforeseen readmissions for new conditions not related to known diseases during the index admissions [28]. The unavoidable readmissions were excluded from the analysis.

From the 13 studies, a total of 42 risk factors were identified and their corresponding predictor variables were aggregated and summarized in Table 2.1. They belong to eight major categories, including sociodemographic factors, healthcare utilization, index admission characteristics, comorbidities and conditions, lab tests, medication, functional status and health literacy, and hospital factors. For each predictor variable, the number of studies found it significant was reported along with the number of studies included it in the analysis. 13 predictor variables were found to be statistically significant ($P < 0.05$) in more than one studies (including age, sex, race, rurality, the insurance payer, the number of hospital admissions in six months or one year before the index admission, the number of emergency department visits in six months or one year before the index admission, the length of stay of the index admission, the type of the index admission, the comorbidity indices, the number of comorbidities, cancer, and the hemoglobin level at discharge). 17 predictor variables were studied in more than one countries or regions and eight of them were found to be significant ($P < 0.05$) in more than

one countries or regions (including age, sex, the number of hospital admissions in six months or one year before the index admission, the number of emergency department visits in six months or one year before the index admission, the length of stay of the index admission, the type of the index admission, the comorbidity indices, and cancer). Of the 42 risk factors, 34 meet our generalizability requirements (with answers NO to the three questions) and were found to be highly generalizable. The corresponding predictor variables of the eight risk factors with low generalizability were labeled with asterisk (*) in Table 2.1 (including the insurance payer, required financial assistance, index admission class, index admission was in a Veterans Affairs hospital, index admission was in a subsidized ward, at-admission activities of daily living, in-hospital activities of daily living decline, and health literacy).

Although it was not the intention of this study to review risk factors only applicable to the United States, about half (7/13) of the studies were based in the United States. To account for the potential bias towards US studies, it is meaningful to compare the risk factors identified within and outside the United States. For each variable, the number of studies found it significant and the number of studies analyzed it were further classified by study regions (either US or non-US) (Table 2.1). For predictor variables only studied in one region, the corresponding numbers in another region were left blank for the sake of clarity. No obvious regional difference was observed for the eight categories, except that the studies in the United States preferred composite comorbidity measures (comorbidity indices and the number of comorbidities) to the presence of individual comorbidities. However, this cannot be justified by significance tests due to the small sample size.

Table 2.1 Summary of corresponding predictor variables of the identified risk factors.

Categories	Predictor Variables	# significant / # analyzed		
		Total	US	Non-US
Sociodemographic factors	Age ^ ~	5/11	2/6	3/5
	Sex ^ ~	2/10	1/6	1/4
	Race or ethnicity ~	2/6	2/4	0/2
	Rurality ~	2/4	2/3	0/1
	Insurance payer *	2/2	2/2	
	Education ~	1/3	0/1	1/2
	Admission class *	1/1		1/1
	Required financial assistance *	1/1		1/1
	Homelessness	1/1	1/1	
	Index admission in a subsidized ward *	1/1		1/1
Healthcare utilization	Number of hospital admissions ^ ~	7/7	3/3	4/4
	Number of emergency department visits ^ ~	2/2		2/2
	Home care services	1/1		1/1
	Nursing home resident	1/1		1/1
Index admission characteristics	Length of stay ^ ~	5/8	4/5	1/3
	Admission type ^ ~	2/2	1/1	1/1
	Admission itself is a readmission	1/1	1/1	
	Discharged from oncology service	1/1	1/1	
	Required inpatient dialysis	1/1		1/1
	Required procedures	1/1	1/1	
Comorbidities & conditions	Comorbidity indices ^ ~	3/5	2/4	1/1
	Number of comorbidities	2/3	2/3	
	Cancer/malignancy ^ ~	2/3		2/3
	Anemia ~	1/2		1/2
	Chronic obstructive pulmonary disease ~	1/2		1/2
	Depression ~	1/2	0/1	1/1
	Diabetes mellitus ~	1/2		1/2
	Heart diseases ~	1/2		1/2
	Acute kidney injury	1/1	1/1	
	Chronic renal failure	1/1		1/1
	Chronic kidney disease	1/1		1/1
	Malnutrition	1/1		1/1
	Sepsis	1/1	1/1	
Lab tests	Hemoglobin level at discharge	2/2	2/2	
	Albumin level ~	1/2	0/1	1/1
	Sodium level at discharge	1/1	1/1	
Medication	Treatment with anti-depressants	1/1		1/1
Functional status & health literacy	At-admission activities of daily living *	1/1		1/1
	In-hospital activities of daily living decline *	1/1		1/1
	Health literacy *	1/1	1/1	
Hospital factors	Bed occupancy	1/1		1/1
	Admitted to a Veterans Affairs hospital *	1/1	1/1	

^ Predictor variables found to be significant in more than one country or region.
~ Predictor variables studied in more than one country or region.
* Predictor variables with low generalizability.

2.4 Discussions

From the 13 studies, 42 risk factors have been identified with 34 being highly generalizable. Their rationale, generalizability, and identification methods will be discussed in this section.

2.4.1 Sociodemographic Factors

In this review, sociodemographic factors were reported by most studies. Age, sex, race, and socioeconomic status are normally used as predictor variables to account for demographic and social influences on readmissions.

Older age has been reported to associate with higher readmission rates [49,63]. The possible reason is that older patients are often frailer and face more health issues than younger patients, such as comorbidities and polypharmacy [49]. Studies have also observed significant differences in readmission rates between sexes [64–66]. Besides biological differences, sex-related social behaviors may play a role in the different readmission patterns [67]. Race and ethnicity can also potentially affect readmissions because they are dimensions of a society's stratification system to distribute resources, risks, and rewards [67].

Socioeconomic status measures an individual or a group's economic and social position by considering income, education, and occupation [68]. Evidence showed that poor physical and psychological health outcomes, including hospital readmissions, were associated with socioeconomic status disadvantage (e.g., low income, limited education, substandard neighborhood) [69–71]. Although the mechanism is still under debate, lower socioeconomic status was reported to indirectly affect health by causing more stress,

exposure to worse physical or social environments, unhealthy lifestyles, or limited access to healthcare resources [72].

Age was considered in 11 studies among which five studies found that increasing age or older age were significantly associated with readmissions. Two studies found that male sex was a risk factor. African American race was found to have a higher readmission risk in two studies. Living in a rural area, having certain insurance payers, education level lower or equal to high school, requiring medical financial assistance are other reported risk factors.

It is worth noting that some factors under this category may depend on or interact with each other. One example is that, in the United States, most people need to reach age 65 to qualify for Medicare, a national insurance program administered by the US government [73]. In this case, Medicare insurance depends on age. These factors can further interact with each other in more implicit ways. Therefore, studies with these factors may need more careful planning and design.

In addition, factors in this category are unmodifiable. It has long been argued in the United States that using readmission rate as a quality indicator without adjusting for unmodifiable socioeconomic factors is unfair because they are beyond the control of hospitals [74]. In 2016, the socioeconomic risk adjustment in hospital readmission measures was finally enforced by the "21st Century Cures Act" [75].

2.4.2 Healthcare Utilization and Medical History

Many studies have incorporated patients' previous healthcare utilization into readmission prediction models. The assumption is that higher utilization such as repeated

admissions to hospitals or emergency departments visits prior to the index admissions may account for the total burden of illness [28], which can potentially relate to readmissions.

Six months or one year are the most common lookback periods to count previous hospital admissions or emergency department visits. A longer look-back period may potentially include utilization less relevant to the readmission of interest and dilute the impact of more recent utilization. Besides higher numbers of previous hospital admissions and emergency department visits, "received home care services" and "being a nursing home resident" were also identified to associate with higher readmission risks.

It is surprising that none of the 13 studies have considered patients' medical history in their analyses. Compared to healthcare utilization, which is high-level information of patients' previous visits, detailed medical history (e.g., diagnoses, procedures, medications, lab test results in visits before the index admission) can provide more information about patients' health status. For example, patients with severe conditions or high-risk surgeries within three months before the index admissions may be at higher risk of unplanned readmission.

2.4.3 Index Admission Characteristics

It has been shown that the length of stay of index admissions can influence readmissions [76]. A longer stay may indicate a more complicated underlying situation and may expose the patient to more risks [45]. However, a shorter stay may also link to a higher readmission risk because the patient may not be ready for early discharge [45,77]. The relationship between the readmission risk and the length of stay has been found to be U-shaped rather than monotonic [76]. In this review, we did not observe a large discrepancy

in the effect of length of stay between the studies as they all agreed that longer index admissions were related to higher risk of readmissions.

Besides the length of stay, the risk factors of acute admission type, admission is a readmission, discharged from oncology service, required inpatient dialysis, and required procedures during the index admission all indicate that patients were in severe situations during the index admissions.

2.4.4 Comorbidities, Conditions, Lab Tests, and Medications

It is well established that comorbidities are associated with undesired healthcare outcomes [78–80]. To date, there has been no consensus on the definition of a comorbidity yet, but the core concept is the coexistence of more than one conditions in the same patient [81]. Evidence shows that the top primary diagnoses of potentially avoidable readmissions are often possible complications of a comorbidity [82] and higher comorbidity has been linked to increased readmission risks [83,84]. In readmission predictions, comorbidities are represented either in the form of the number of comorbidities, the comorbidity index, or the presence of a comorbid condition.

It has been found that the readmission risk will rise as the number of comorbidities increases from the reviewed studies. More than just counting the number of comorbidities, the comorbidity index further accounts for contributions of different comorbidities. Charlson [85] and Elixhauser [86] are the most commonly used comorbidity indices [87]. The Charlson index was originally developed based on medical record review of 19 comorbid conditions with each condition assigned a weight of 1, 2, 3, or 6 depending on the risk associated with mortality [85]. A higher total index indicates a greater chance of

one-year mortality. The Charlson/Deyo index is a highly referred variant by adapting the original index to 17 categories of comorbid conditions with International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) codes [88]. The Exlihauser index includes a more comprehensive list of 30 comorbidities [86] but with little overlap with the Charlson index [87]. According to a systematic review of 54 articles in 2012, the Elixhauser index generally outperforms other available indices [87].

The presences of some chronic or acute conditions are also related to readmissions. Especially, cancer, chronic obstructive pulmonary disease, heart diseases, renal diseases, diabetes mellitus, and sepsis found in this review are among conditions associated with the most readmissions [35]. The included lab tests and medication are closely related to some conditions on the list, such as Anemia, renal diseases, and depression.

2.4.5 Functional Status and Health Literacy

According to Leidy's definition, functional status measures a person's ability to provide for the necessities of life, including daily activities to meet basic needs, fulfill usual roles, and maintain health and well-being [89]. The impairment of functional status has been reported to associate with increased risk of readmissions [90].

Health literacy was defined as "the degree to which individuals have the capacity to obtain, process, and understand basic health information and services needed to make appropriate health decisions" by Ratzan et al. in 2000 [91] and this definition was adopted by the Institute of Medicine of the United States [92]. Although not considered as a social factor, it is more distally influenced by social factors [67]. Low health literacy may attribute to nonadherence to treatment plans, compromised communications with clinicians, limited

self-care skills [93], and is associated with many poor health outcomes, including hospital readmissions [94]. To assess health literacy, questionnaire-based tests are administered and several tools are available [95].

Evidence showed the inclusion of functional status or health literacy can increase the predictive performance of readmission models [52]. However, they are seldom used due to the difficulty of data collection [28], especially in the case of retrospective studies.

2.4.6 Hospital Factors

The factors from the hospital side may also contribute to readmissions in many ways. For example, the pressures in hospital resources (e.g., beds) may cause premature discharges of existing patients, which have shown to be related to readmissions [77]. There is also evidence that medical errors associate with higher readmission risks [96].

However, similar to the finding of another study [50], most of the identified risk factors are patient-side factors or clinical factors and only two hospital-side risk factors (inpatient bed occupancy > 95%, index admission was in a Veterans Affairs Hospital) were found. This could be attributed to the small sample size, but a more plausible reason is that most studies followed the single-center retrospective cohort design. For single-center retrospective studies, it is harder to collect encounter-level hospital-side factors. If possible, it is recommended to collect multi-center data or combine with other data sources, such as claims data, to account for the variances in hospital-side factors.

Another possible reason is that, unlike patient-side factors and clinical factors, which are usually well-defined and readily available in administrative and clinical databases, hospital-side factors are harder to collect. More efforts are needed to define and

quantify hospital-side factors in higher granularity beyond the basic hospital characteristics, such as geolocation, hospital type, teaching status, and beds, especially for studies measuring readmission rates for quality compare purposes.

2.4.7 Generalizability of the Risk Factors

The objective of this study was to review risk factors that can be widely generalized regardless of target populations and their residing countries. The generalizability of the 42 identified risk factors was assessed by the three questions detailed in the methods.

34 of the risk factors meet our generalizability requirements (with answers of NO to the three questions). The corresponding predictor variables of the eight risk factors with low generalizability include the insurance payer, required financial assistance, index admission class, index admission was in a Veterans Affairs hospital, index admission was in a subsidized ward, at-admission activities of daily living, in-hospital activities of daily living decline, and health literacy.

Health insurance is country specific. “Medicare/Medicaid as insurance” and “Medi-Cal as insurance” are significant but they are only applicable in the United States. The insurance payer is a useless predictor variable for countries with universal healthcare coverage. “Requiring financial assistance from Medifund”, “index admission was in a subsidized ward”, and “index admission class > A” may indicate a lower socioeconomic status but they all closely relate to the financial regulations and social welfare of the patients’ residing countries. The difficulty of collecting these data can be distinct in different countries. “Index admission was in a Veterans Affairs hospital” is only valid in

the United States. We excluded functional status and health literacy because they are often harder to collect (e.g., interviews, self-reporting) for retrospective studies.

We kept comorbidity indices in the list of highly generalizable risk factors. Although the Charlson/Deyo and Elixhauser indices were originally built based on ICD-9-CM codes, which were the adaption of ICD-9 codes in the United States [97], they have been successfully translated to work with ICD-10 codes in Canada and Switzerland [62,98].

2.4.8 Timeliness of Variables

Some studies developed readmission predictive models with variables whose values can only become available near or after discharge, such as length of stay, discharge disposition, and lab test results before discharge. At the end of inpatient care, patients' clinical information tends to be completer and more accurate than at the early stage of care. Including these variables can potentially improve the predictive performance. However, this will limit the timeliness of the predictive model. As a result, models with these variables cannot be used for early prediction of readmission.

It has been argued that inclusion of comorbidity measures or diagnosis codes in readmission prediction models may reduce the timeliness of the predictions. The reason is that in EHR, this information is represented in ICD codes, which are retrospectively assigned by medical coders near or after the end of inpatient encounters for billing purpose [28]. Nevertheless, in practice, clinicians will have this information during the inpatient care based on their medical judgements without the help of coders.

2.4.9 Methods to Identify Risk Factors

The reviewed studies are highly consistent in analytical methods. 12 studies used logistic regression and one used Poisson regression. Logistic regression and Poisson regression both belong to the family of generalized linear models, which estimate model parameters by maximizing likelihood [99]. Poisson regression assumes the response variable follows a Poisson distribution, while in logistic regression the response variable can be either binomial, ordinal, or multinomial. Binomial logistic regression is usually used in readmission predictions because the outcome is dichotomous (either readmitted or not readmitted). In binomial logistic regression, the binary response variable is linked to the linear combination of independent predictor variables through a logit function [100]. Poisson regression models a discrete count response variable with the logarithm as the link function [99]. In these studies, the adjusted odds ratio (OR) was the most used metric to assess a variable's degree of association to the response variable. The odds ratio measures the relative chance of an outcome of interest to occur under different exposures [101]. The significance levels were set to 0.05 in all the studies.

Surprisingly, none of the 13 reviewed studies has used methods other than traditional statistical analysis. In recent years, data mining has been a hot research area and there have been many successful applications in healthcare [102]. Unlike statistics, which is hypothetico-deductive, data mining uses more flexible and more inductive ways to find patterns hidden in data [102]. Decision trees [103] are a family of supervised classifiers especially suitable to identify risk factors. The process to assign a label to the response variable can be visualized in a straightforward tree-like structure. The critical cutoff values of predictor variables associated with readmissions can be directly obtained from decision

trees. The association rule mining [104] is another data mining technique appropriate to identify risk factors. This technique intends to discover strong rules (frequent item sets) based on predefined criteria. Risk factors can be extracted from the rules with high ranks.

Another concern is that some studies reported results without evaluations and/or internal validations of the prediction models. To reduce the bias and improve the usefulness of a prediction model, it is recommended to report prediction models following the guidelines in the TRIPOD statement [105] and the statement from the American Heart Association [106]. Especially, it is important to evaluate and report the model's performance in the derivation and validation datasets.

The most popular model evaluation metric is the area under the receiver operating characteristic curve (AUC), or called the c-statistic [107]. The receiver operating characteristic curve (ROC) is a graphical representation of a binary classifier's performance as the discrimination threshold is varied [108]. The AUC measures the model's ability of discrimination and can be interpreted as the probability that the model will rank a randomly selected positive sample higher than a randomly selected negative sample [108]. The AUC ranges from 0.5 to 1 with 1 indicating a perfect classifier.

The two widely used validation methods are hold-out cross-validation and k-fold cross-validation [109]. The hold-out method splits the dataset into a derivation dataset and a validation dataset. The derivation dataset is used to build the prediction model and the validation dataset is used to test the model. The disadvantage of the hold-out method is the partition of the original dataset might be biased and the resulting derivation and validation datasets might follow different local distributions. To overcome this issue, k-fold cross-

validation method randomly splits the original dataset into k equal-sized partitions and uses one partition as the validation dataset and the remaining partitions as the derivation dataset. This process will be repeated k times and the k validation results will be averaged as the final validation result.

2.5 Limitations

This study has a couple of limitations. First, due to the strict inclusion criteria, only 13 articles were selected into the final literature review and 15/34 of the highly generalizable risk factors were reported in only one study. Because of the small sample size, it is infeasible to conduct statistical significance tests. However, the intent of this study was not to review risk factors that shared by most studies. Instead, the objective was to provide a list of highly generalizable risk factors to guide the selection of baseline predictor variables in different readmission studies. Even if some risk factors were reported by only one study, we chose to keep them because they were reported to be statistically significant in the prediction of readmissions and can be easily applied to other studies.

Second, the articles are imbalanced in study regions with about half (7/13) based in the United States. This may introduce bias and potentially weaken the generalizability of some risk factors. However, after comparing the US and non-US studies, we did not find an apparent difference in most risk factor categories. Studies in the United States are more likely to use composite comorbidity measures such as comorbidity indices and the number of comorbidities other than individual comorbidities. Although the reported comorbidity indices were originally developed in the United States based on ICD-9-CM codes, they have been translated to work with ICD-10 codes and have been applied globally.

2.6 Conclusions

In this work, we have identified 34 highly generalizable risk factors for unplanned 30-day all-cause hospital readmissions. They are not specific to any populations or places and the corresponding predictor variables can potentially serve as baseline predictor variables in readmission prediction studies around the world. The majority of the identified risk factors are patient-side factors and clinical factors. Only two hospital-side factors have been identified. This could be due to the limitation of the study design and the difficulty of data collection. Compared to healthcare utilization, the impact of detailed medical history on readmission is under-explored. Some models do not support the early prediction of readmission because they used variables whose values can only become available near or after discharge. No major difference has been observed between the risk factors identified inside and outside the United States except that US studies appeared to prefer composite comorbidity measures. However, this assertion should be validated by significance tests when more eligible studies become available. All the reviewed studies have used traditional statistical regression-based methods to identify risk factors. More applications of modern data mining techniques in readmission prediction studies are expected. Overall, the literature suggests a growing interest in developing hospital readmission models in the past decade. The findings of this review can guide the selection of baseline readmission predictor variables and potentially provide the foundation for international collaborations on readmission predictions.

Chapter 3 An Early Prediction Model of Unplanned 30-Day Hospital Readmission

3.1 Background

Predictive modeling is an efficient way to reduce unplanned 30-day hospital readmission because it can stratify patients' readmission risk and target preventive interventions to patients at high risk [22]. Evidence has shown that applying interventions to high-risk patients can reduce 30-day hospital readmission risk by 11-28% [23–25]. However, the majority of reported hospital readmission predictive models have limited clinical values because they are based on administrative data or require variables whose values only become completely available at discharge, such as the length of stay, lab test results before discharge [26]. For example, the HOSPITAL score [28] and the LACE index [27] are the most widely used readmission risk calculators in the US healthcare settings, and both of them can only be used near discharge. After patients have been discharged, clinicians can no longer provide impactful care. It has been shown that early interventions, such as early discharge planning [21] can reduce readmissions. It is essential to perform early risk assessment of high-risk patients to deliver effective preventive interventions during hospitalization instead of near or after discharge [29].

Several early prediction models of readmission have been reported but their performance and design are unsatisfactory. Wang et al. [110] developed a real-time model using the time series of vital signs and discrete features, such as lab tests. However, this

model was based on deep neural networks and cannot be interpreted. In healthcare applications, a model's interpretability is as important as the performance because the attributes and the decision path need to be medically rational. Horne et al. [111] developed a laboratory-based model specific to heart failure patients. It can be used within 24 hours of admission, but the performance was poor with the AUC of 0.571 and 0.596 in female and male validation sets. Cronin et al. [112] reported an early detection model based on the information available at admission and index admission's medication record with a moderate performance (AUC, 0.671) on the validation set. El Morr et al. [113] created a modified LACE index (LACE-rt) to support real-time prediction by replacing the length of stay during the current admission with that of the previous admission within last 30 days. However, it has a fair performance (AUC, 0.632) [113]. In addition, none of these studies excluded planned readmissions following the CMS guideline [33].

The aim of this chapter was to build a predictive model for early detection of readmission with statistical and machine learning methods. Based on the systematic review of readmission risk factors in Chapter 2, we found that medical history was under-explored in other studies. In this chapter, we included the detailed medical history of previous encounters up to one year before index admissions in the readmission prediction model.

3.2 Materials and Methods

3.2.1 Study Design

This study was a retrospective analysis of EHR data. To ensure that our readmission prediction model can work at the early stage of hospitalization, we only used index admission's attributes whose values can be available in EHR within 24 hours, including

patients' demographics, lab tests, vital signs, as well as medications. Patients' data was enriched by the detailed medical history of previous hospital encounters within one year before the current inpatient stay, including the information of diagnosis, procedure, lab test, vital sign, medication, and healthcare utilization. Figure 3.1 shows the design.

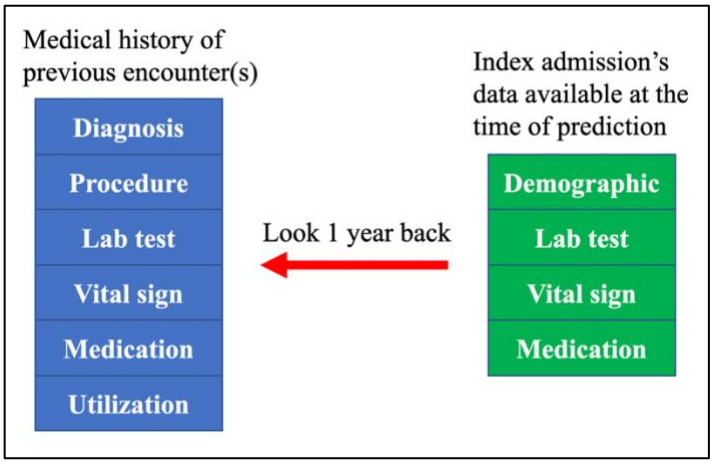


Figure 3.1 Variables of the early prediction model.

3.2.2 Data Source

The data was extracted from Health Facts® [114], a de-identified EHR database curated by Cerner Corporation and hosted by the School of Medicine at the University of Missouri. As of 2018, it contains 3.15 TB de-identified EHR-level data from 782 Cerner client hospitals and clinics in the United States between 2000 and 2016.

3.2.3 Data Inclusion and Exclusion Criteria

The data inclusion and exclusion criteria were based on the criteria used by the CMS [33] with minor modifications. (1) We captured inpatient encounters between Jan 1st, 2016 and Dec 31st, 2016 in acute care hospitals with a length of stay longer than one day. Patients cannot be readmitted for rehabilitation services and the gap between index

admission discharge and readmission is between one and 30 days (inclusive). If a patient had more than one inpatient visit within 30 days of discharge, only the first one will be considered as readmission. (2) Patients were older than 18 at admission. They were not transferred to other acute care facilities and were alive at discharge. (3) Patients were not readmitted for newborn, labor, accident, trauma, or other scheduled care according to the CMS' planned readmission identification criteria [33]. In this work, we adopted the concept of "hospital-wide all-cause readmission" used by the CMS [115] because we are interested in readmissions caused by medical and healthcare-related reasons. (4) We used the same criteria of index admissions to identify control patients who did not experience readmission. (5) Each patient has only one index admission during a readmission episode.

3.2.4 Feature Engineering

According to the systematic review of readmission risk factors in Chapter 2, patients' demographic and social factors, as well as previous healthcare utilization are strong predictors for readmission. In this work, we incorporated patients' age at admission, sex, race, the insurance payer, the hospital's census region, census division, rurality, and healthcare utilization in the previous year, including the number of inpatient visits, the number of outpatient visits, the number of emergency department visits, and the number of leaving against medical advice. We also investigated the impact of medical history within a year before the index admission. We used counts to condense the longitudinal medical history into structured data so that patients with different medical histories can be represented in the same feature space. Patients with more previous visits will have higher counts and patients without any medical history will have counts of zero. In this way, we will be able to handle the missing value problem for new patients. For example, if a patient

has the same diagnosis of heart failure in two separate encounters last year, this diagnosis will have a count of two. For lab tests and vital signs, only the latest result will be checked to see if it was abnormal or not. Suppose a patient took the systolic blood pressure twice in one encounter and the result of the latter test was abnormal. In another visit, he took it three times and the latest result was normal. Then, this patient will have one abnormal systolic blood pressure during the two encounters in last year. For the index admission, we only checked the medication record and the latest result of lab tests and vital signs. Diagnosis codes were mapped from International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM) code [116] into the CCS categories [117] because ICD codes were too granular for data mining purposes. For the same reason, procedure codes were mapped from International Classification of Diseases, Tenth Revision, Procedure Coding System (ICD-10-PCS) [118], Current Procedural Terminology (CPT) [119], and Healthcare Common Procedure Coding System (HCPCS) [120] codes into CCS categories. Lab tests and vital signs were represented by their original names. We used generic names to represent medications without grouping. Table 3.1 shows information of these features.

Table 3.1 Feature representation and value type.

Type	Category	Representation	Data Type
Medical History in Last Year	Diagnosis	CCS	Count
	Procedure	CCS	Count
	Lab test	Name	Count
	Vital sign	Name	Count
	Medication	Generic name	Count
	Utilization	Name	Count
Index Admission	Demographic	Name	Discretized age, race, sex, payer, region, rurality
	Medication	Generic name	Boolean - The medication is taken or not
	Lab test	Name	Boolean - The latest result is abnormal or not
	Vital sign	Name	Boolean - The latest result is abnormal or not

3.2.5 Algorithms and Models

We selected six candidate algorithms that can generate probabilistic outputs, including logistic regression, naïve Bayes, decision trees, random forests, gradient tree boosting, and artificial neural networks. Logistic regression belongs to the generalized linear models [99] family and predicts the log odds of the positive response as a linear combination of variables weighted by coefficients [100]. The contribution of a variable (factor) to the prediction can be measured by the odds ratio (OR) [121], which equals to the exponential of the variable's coefficient. An odds ratio greater than one indicates the corresponding factor is a risk factor whose presence raises the odds of the positive outcome (e.g., readmission). An odds ratio lower than one indicates a protective factor whose presence reduces the odds of the positive outcome. Naïve Bayes is a probabilistic classification algorithm based on Bayes' theorem [122] with the assumption that variables are independent [123]. Classifications are achieved by assigning the class label that can maximize the posterior probability given the features of an instance. Naïve Bayes model can be interpreted by taking the conditional probability of a variable given a class, and a higher probability indicates a stronger relationship with the class. Decision trees are a family of tree-structured predictive algorithms, which iteratively split the data into disjoint subsets in a greedy manner [124]. Classifications are made by walking the tree splits until arriving a leaf node (the class). Decision trees are self-explainable because each leaf node is represented as an if-then rule and the decision process can be visualized. The contribution of a variable to the classification can be measured using various methods such as information gain based on the information theory and Gini importance [125]. Random forests are an ensemble learning algorithm generated by bootstrap aggregation, which

repeatedly selects a random sample from the training set (with replacement) and builds a decision tree for the sample [126]. When making predictions, the outputs from different decision trees will be ensembled. Gradient tree boosting is another type of tree ensemble algorithm, which builds the model in a stage-wise fashion by iteratively generating new trees to improve the previous weaker trees [127]. Predictions are made by weighted average of tree outcomes with stronger trees having higher weights. Random forests and gradient tree boosting algorithms can be interpreted by measuring variables' Gini importance. Artificial neural networks are an interconnected group of computing units called artificial neurons [128]. Artificial neurons are aggregated into layers and connected by edges, which have different weights to control the signal transmitted between neurons. The signals in the final output layer are used for prediction. Each feature's importance can be measured by the increase in prediction error after permuting the feature's values.

We implemented the HOSPITAL score, the LACE index, and the LACE-rt index so that we can compare their performance with our models. Appendix 1 shows their point systems and Python implementations. The HOSPITAL score has seven variables, including hemoglobin level at discharge, discharge from an oncology service, sodium level at discharge, any ICD procedures during the hospital stay, the type of index admission, the number of admissions one year before the index admission, and length of stay [28]. Each factor level has a weighted point and the total score can be up to 13 points. The LACE index has four variables, including length of stay, acuity of admission, the Charlson comorbidity index, and the number of emergency department visits six months before the index admission [27]. It ranges from 0 to 19 points. The LACE-rt index has the same variable weights and the same maximum score as the original LACE index. The only

difference is that it requires the length of stay during the previous admission within last 30 days instead of the current admission.

To evaluate the models' performance, we used AUC, precision, recall, specificity, and F1-measure as metrics. AUC is the probability that the model will rank a randomly chosen positive instance higher than a randomly chosen negative instance. AUC ranges from 0.5 to 1.0 with 1.0 indicating the model has a perfect discrimination ability and 0.5 meaning it is no better than random guess. Precision is the fraction of true positives among all instances predicted to be positive. Recall is the fraction of correctly identified positives in all positive instances. Specificity is the fraction of correctly identified negatives in all negative instances. F1-measure is the harmonic mean of precision and recall. The values of precision, recall, specificity, and F1-measure are between 0 and 1.0. A higher value indicates better performance.

3.3 Results

After data transformation and feature engineering, the final data set has 96,550 records and 432 variables. The readmission rate (11.7%) is lower than the Medicare readmission rate (15.2% [2]) because we included patients between 18 and 64 years old, who were younger and less vulnerable. Table 3.2 shows the demographic information of the 96,550 patients. The majority of patients (71.9%) are Caucasian. 26.9% of patients are between 65 to 79 years old. More than half of patients (57.6%) are female.

Table 3.2 Demographics information of the 96,550 included patients.

Factor	Frequency	Percentage	Readmission Rate
Age			
18-34	14,172	14.7%	6.6%
35-49	14,066	14.6%	10.8%
50-64	24,675	25.6%	12.6%
65-79	**26,014**	**26.9%**	13.0%
80+	17,623	18.3%	**13.3%**
Sex			
Female	**55,585**	**57.6%**	10.7%
Male	40,965	42.4%	**13.0%**
Race			
African American	18,860	19.5%	**13.8%**
Caucasian	**69,435**	**71.9%**	11.2%
Other	8,255	8.5%	11.3%

We randomly split the 96,550 records into a development set (91,550 records) and a validation set (5,000 records). The readmission rate (11.7%) was preserved in these two data sets. The development set was used to derive and test the five candidate models in 10-fold cross-validation. The validation set was kept intact until the last moment to assess the models' generalizability on unseen data. We extracted variables required by the HOSPITAL score, the LACE index, and the LACE-rt index from encounters in the validation set to test their performance. Table 3.3 shows the AUC of the models on the development set (10-fold cross-validation). Especially, the alternating decision tree (ADTree) [129], the extreme gradient boosting (XGBoost) [130] algorithms, and the feedforward neural networks with three hidden layers (256 neurons, 512 neurons, and 256 neurons) had the best AUC within the decision trees, gradient tree boosting, and artificial neural networks families, respectively. The XGBoost model achieved the overall best AUC

of 0.753 on the development set. Figure 3.2 shows the ROC curves of the 10-fold cross-validation of the XGBoost model.

Table 3.3 AUC of the six candidate models on the development set.

Model	AUC
Logistic Regression	0.750
Naïve Bayes	0.730
ADTree	0.730
Random Forests	0.733
XGBoost	**0.753**
Neural Networks	0.746

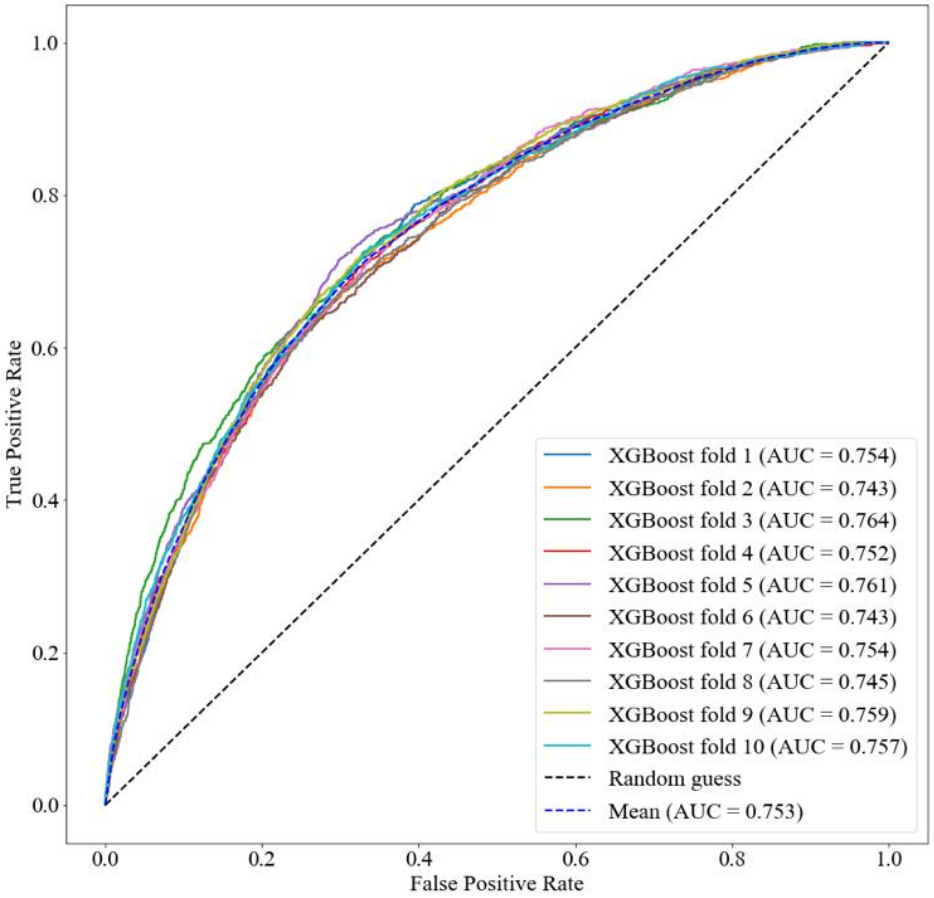


Figure 3.2 ROC curves of 10-fold cross-validation (XGBoost).

We further compared the models' precision, recall, specificity, F1-measure, AUC, and optimal cutoffs on the validation set in Table 3.4. Because the prevalence of readmissions is imbalanced in nature (e.g., 11.7% in this study), it is infeasible to use 0.5 as the cutoff probability to dichotomize probabilistic outputs. We chose cutoffs that can maximize each model's Youden's index [131], which equals to the sum of recall and specificity minus 1. The cutoffs of the three baseline models are scores because they do not generate probabilities. It can be seen that the random forests model has the best specificity and precision, while the XGBoost model has the best recall, F1-measure, and AUC. In medical domain, recall is a more important metric because false negatives are considered more expensive than false positives. Based on the recall and AUC, we chose the XGBoost model as the final model. For all performance metrics, the XGBoost model is better than the three baseline models. Figure 3.3 shows the ROC curves of the XGBoost model and the three baseline models on the validation set. In addition, the importance of features of the XGBoost model is shown in Appendix 2 – Appendix 4.

Table 3.4 Comparison of models on the validation set.

Model	Optimal Cutoff	Specificity	Precision	Recall	F1-Measure	AUC
Logistic Regression	0.157	0.642	0.857	0.729	0.773	0.741
Naïve Bayes	0.220	0.666	0.855	0.685	0.740	0.720
ADTree	0.298	0.662	0.857	0.705	0.755	0.732
Random Forests	0.122	**0.747**	**0.862**	0.611	0.680	0.726
XGBoost	0.175	0.611	0.856	**0.759**	**0.794**	**0.743**
Neural Networks	0.125	0.686	0.858	0.681	0.737	0.735
HOSPITAL score	4	0.564	0.838	0.694	0.745	0.688
LACE index	11	0.469	0.830	0.745	0.779	0.675
LACE-rt index	7	0.542	0.833	0.688	0.740	0.668

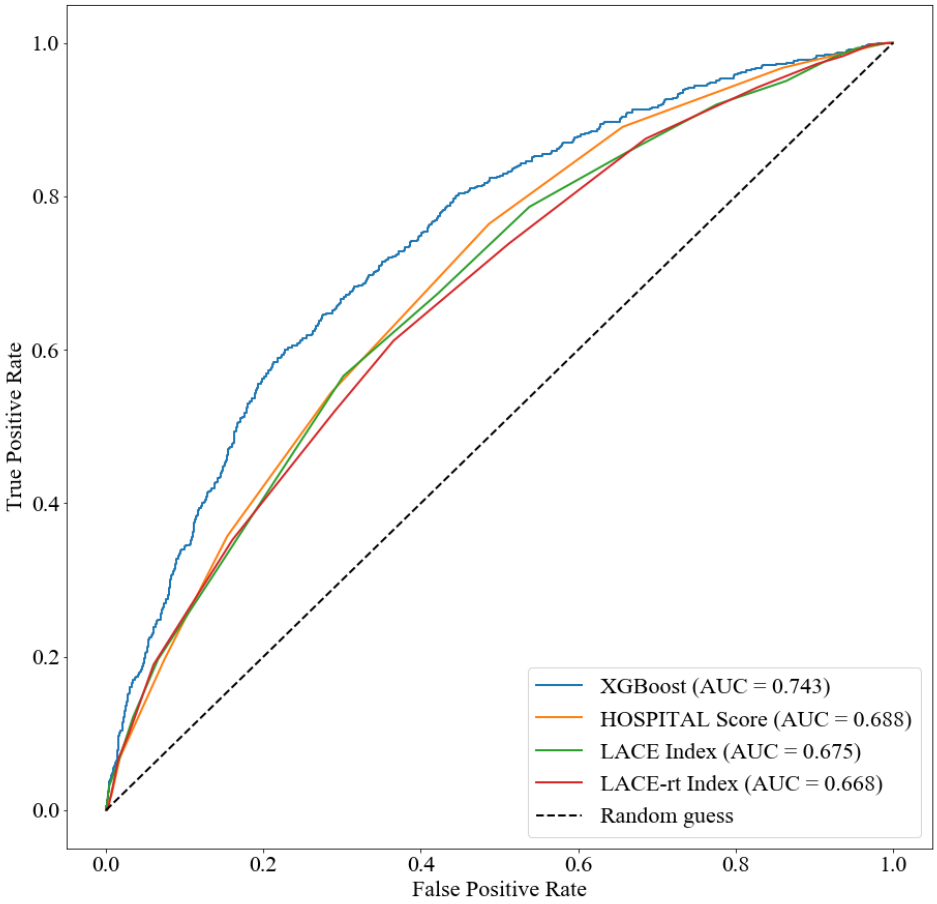


Figure 3.3 ROC curves of the XGBoost and baseline models on the validation set.

To better understand the statistical significance of factors, we performed the multivariate analysis on the whole data (96,550 records and 432 variables). By backward elimination, we reduced the feature space down to 83. We re-identified 40 risk factors and significant predictors reported by other studies before. In addition, we discovered 14 novel risk factors and two novel protective factors that have never been reported in the literature. They belong to 13 predictor variables and are displayed in Table 3.5.

Table 3.5 14 novel risk factors and two novel protective factors for readmission.

Risk/Protective Factors	Coefficient	P-value	OR	95% CI
# maintenance chemotherapies last year				
1	0.390	< 0.001	1.476	1.218 - 1.790
# abnormal lymphocyte count tests last year				
1	0.221	< 0.001	1.247	1.144 - 1.359
≥ 2	0.228	0.001	1.257	1.091 - 1.447
# abnormal monocyte count tests last year				
1	0.182	0.005	1.199	1.056 - 1.362
# abnormal monocyte percent tests last year				
≥ 2	0.316	< 0.001	1.371	1.178 - 1.596
# abnormal serum calcium quantitative tests last year				
1	0.226	< 0.001	1.254	1.107 - 1.420
≥ 2	0.297	0.001	1.345	1.122 - 1.612
# prescriptions of albuterol ipratropium last year				
1	0.071	0.023	1.073	1.010 - 1.141
≥ 2	0.145	0.003	1.157	1.052 - 1.272
# prescriptions of cefazolin last year				
1	-0.123	0.001	0.884	0.822 - 0.950
Index admission hospital census region				
Northeast	0.365	< 0.001	1.441	1.345 - 1.543
Prescribed gabapentin in index admission				
Yes	0.162	< 0.001	1.176	1.113 - 1.243
Prescribed ondansetron in index admission				
Yes	0.105	< 0.001	1.111	1.057 - 1.168
Prescribed polyethylene glycol 3350 in index admission				
Yes	0.073	0.011	1.076	1.017 - 1.139
Prescribed cefazolin in index admission				
Yes	-0.147	< 0.001	0.863	0.798 - 0.934
# abnormal lab tests in index admission				
≥ 16	0.140	0.005	1.151	1.043 - 1.269

3.4 Discussions

In this work, we developed an early prediction model of hospital readmission. The XGBoost model has the best recall, F1-measure, and AUC. Using multivariate analysis, we identified 14 novel risk factors and two novel protective factors for readmission.

3.4.1 Novel Risk Factors and Protective Factors for Readmission

The 14 novel risk factors and two protective factors for readmission are related to medical history and index admission. They can be classified into four categories, including diagnosis, hematology and blood chemistry tests, medications, and the census region.

Patients with one CCS-level diagnosis of maintenance chemotherapy in the previous year was found to be more associated with readmission than patients without it. This can be explained by the linkage between chemotherapy and cancer, which has been reported as a predictor of readmission [52,55].

Blood disorder or an abnormal amount of substance in the blood can indicate certain diseases or side effects. Having an increased number of abnormal test results indicates that the patient is frailer and can be more prone to readmission.

The prescriptions of four medications were found to be positively linked to readmission. These medications may have side effects that are associated with readmission. Another interpretation is that conditions treated by these medications may be related to readmission. For example, albuterol ipratropium is a combination of two bronchodilators, which are used in the treatment of COPD. COPD has been reported as a risk factor of readmission [55]. It is interesting that the prescriptions of cefazolin in previous encounters and index admission were both negatively associated with readmission (protective factors). One possible explanation is that cefazolin is a type of antibiotics, which are used to treat infections caused by bacteria. The usage of cefazolin may potentially reduce patients' chance of infection and reduce the readmission risk.

The Northeast census region was found to be more positively associated with readmission than Midwest census region. One possible reason is that geo-location is associated with socioeconomic status, which has been reported to be linked to readmission [59].

3.4.2 Timeliness of Prediction

Most readmission predictive models are based on index admission's data. Many highly predictive variables of the index admission, such as the length of stay, discharge disposition, and lab test results before discharge are only available near or after discharge. To achieve good predictive performance, most studies include these variables in their models. As a result, these models can only be used near or after discharge. They are good for public reporting but not clinical decision support because they are not timely.

In this work, we used the data of index admission and patients' medical history up to one year before the index admission. To ensure the model can work in the early stage of the hospitalization, we only used index admissions' data that can become available within 24 hours in EHR during hospitalization, such as medication, lab tests. We used the detailed medical history of previous encounters. Although healthcare utilizations have been used in other studies, they are only high-level information of previous encounters (e.g., the number of inpatient stays last year) instead of detailed information, such as previous lab test results. By using the detailed medical history, we were able to add more variables to the model without sacrificing its timeliness. As a result, our model enables point-of-care prediction and can be used to continuously monitor the readmission risk during the whole episode of hospitalization.

3.4.3 Generalizability

Besides performance and interpretability, we also considered the model's generalizability. From the modeling point of view, generalizability indicates if a model can achieve similar performance on data that has never been seen by the model. In other words, the model should be trained and built using a large and diverse training sample to represent the whole population. Most existing readmission prediction models were based on relatively homogenous (e.g., single-center studies) and small (e.g., less than 20,000) samples. For example, the HOSPITAL score and the LACE index were derived from only 9,212 American patients and 4,812 Canadian patients, respectively [27,28]. To ensure good generalizability, we captured all eligible readmissions in 2016 from the Health Facts® database and the final data contained 96,550 patients discharged from 205 hospitals across the four US census regions. For the best performing candidate model (XGBoost), the AUC on the validation set is close to the AUC on the development set (0.743 versus 0.753), which indicates that the model has good generalizability. Furthermore, our model is better than the three baseline models for all of the five performance metrics on the validation set.

Another consideration of generalizability is if the model can work on various types of patients. There is no consensus on data inclusion criteria for readmission studies and the study outcomes span from condition-specific to all-condition readmission predictive models [34]. The choice between these two types of models has long been under debate. In two systematic reviews [22,26] of 99 readmission predictive models reported between 1985 and 2015, 77% of the models are specialized for one patient subpopulation. The condition-specific design limits the models' adaptability in other patient subpopulations and may potentially overlook some at-risk minorities if specific models are not available

for them [36,37]. In practice, it can be challenging for a hospital to maintain separate readmission prediction models for different patient subpopulations, and this situation will be further exacerbated if patients have comorbidities [37]. All-condition models are designed for broad patient populations without limiting diagnoses or procedures. In this work, we are interested in hospital-wide readmissions caused by medical and healthcare-related reasons. Our model is not specific to any conditions or procedures because we want to use it as an early screening tool to assess all patients' risk.

3.5 Limitations

Although our model was designed to be unspecific to patient populations, it does not work for patients under 18 years old. The reason is that the infant and pediatric readmissions were reported to have different patterns from adult readmissions [26,50] and could be influenced by parental factors [50,51]. The Health Facts database is de-identified and there is no information about a patient's family. Therefore, we removed patients younger than 18 from the data. Besides, the Health Facts database only contains data collected from US healthcare settings. For readmissions in other countries, whose patients' demographics and medical interventions (e.g., race, medications) are different from the United States, our model may not work well.

3.6 Conclusions

While the whole healthcare industry is focusing on improving the transitions of care and post-discharge care, interventions during hospitalization hold the potential for reducing readmission risk. It is challenging for clinicians to identify patients with high risk for readmission at the early stage of hospitalization because little data is available. In this work,

we have developed an early prediction model for unplanned 30-day hospital readmission. Our model uses the detailed information of patients' index admissions and medical history up to one year prior to the index admissions. Unlike most models, which can only make predictions near or after discharge, our model can monitor patients' readmission risk at the beginning of care. This feature allows clinicians to design and deliver interventions to mitigate the readmission risk before patients are discharged. Compared to most existing readmission prediction models, our model was derived and validated from a larger and more diverse patient population (96,550 patients discharged from 205 hospitals across four US census regions). This ensures that our model can generalize well to adult patients in the United States. The predictive performance of our model is better than the HOSPITAL score, the LACE index, and the LACE-rt index on the validation data. By multivariate analysis, we identified 14 novel risk factors and two novel protective factors of readmission. To our knowledge, they have never been reported. This may shed a light on the understanding of the complex readmission problem, but more studies or trials are necessary to verify these predictors' relationship with readmission.

Chapter 4 Identification of Potentially Modifiable Risk Factors for Unplanned 30-Day Hospital Readmission

4.1 Background

Existing unplanned hospital readmission reduction programs tend to focus on care transition and post-discharge interventions. However, one limitation of this approach is that they are mostly initiated near or after discharge when clinicians are no longer impactful on inpatient care. Preventive intervention during hospitalization is an under-explored area that holds the potential for reducing readmission risk. However, there are two challenges to deliver interventions during hospitalization. First, it is difficult to foresee the causes of readmissions because patients can be readmitted for any reasons. The CMS uses index admissions' principal diagnoses and procedures to define the six HRRP eligible cohorts but imposes no restriction on the causes of readmissions. Many patients are readmitted for different reasons. According to an analysis of 217,767 index admissions with readmissions by Rosen et al. [38], about 60% of readmissions have different principal diagnoses, different DRG, or different procedures from the precedent index admissions. Second, it is challenging for clinicians to find a universal solution to reduce their readmission risk because patients are heterogeneous. Even if they share the same principal diagnoses and procedures in index admissions, they may have different combinations of comorbidities and social factors. It has been found that direct or indirect complications of patients' comorbidities are the top causes of readmissions [82]. Social factors, such as demographics

and socioeconomic status, can also influence readmission risk [49,71]. Therefore, interventions should be customized based on patients' specific clinical and social factors.

To address the above two challenges, we purposed to identify patterns with potentially modifiable risk factors and recommend them to different patient subgroups to support the development of customized interventions. Potentially modifiable risk factors are risk factors that can be potentially controlled or treated by interventions [132]. It has been shown that risk factor modification was effective in reducing the risk of other undesired outcomes [133–136]. To identify potentially modifiable risk factors, we investigated a novel method to compare different index admissions of the same patients. This method allows us to explore associations between changes of modifiable risk factors and the change of readmission status. We focused on medical services for AMI, COPD, HF, and PN because they are targeted by the HRRP. Surgical services were not included in this analysis because surgical readmissions tend to be related to postoperative complications [38], which are undesired outcomes of medical care. Surgical readmissions can be potentially reduced by minimizing the risk of complications during the care process.

4.2 Materials and Methods

4.2.1 Study Design

This study was a retrospective analysis of clinical data. We identified pairs of index admissions of the same patients and recorded the element-wise difference of each diagnosis' existence and the difference of readmission status in each pair. Figure 4.1 shows an example of this comparison. Suppose that a patient has a pair of index admissions 1 and 2. This patient has diagnosis A in both index admissions and diagnosis B in index admission

1 only. Index admission 1 was followed with a readmission, and index admission 2 was not. The change of diagnoses' existence in these two index admissions was "diagnosis B: presence (T) → absence (F)" and the change of readmission status was "readmission → no readmission". Each patient has one pair of index admissions. We looked for associations between changes of diagnoses' existence and the change of readmission status.

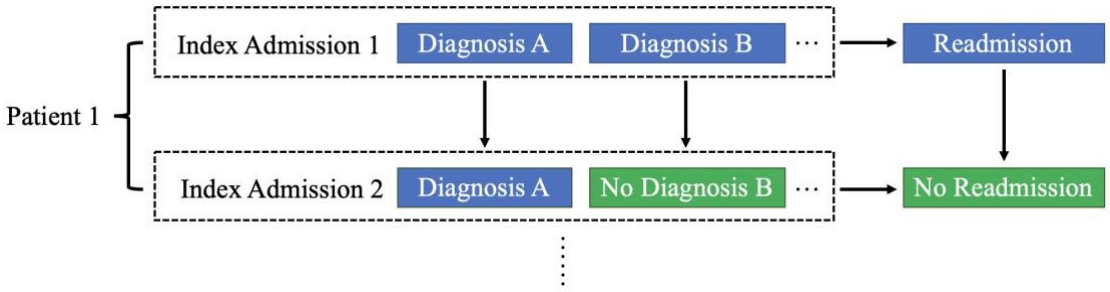


Figure 4.1 An example of the elementwise comparison between two index admissions of the same patient.

4.2.2 Data Source and Inclusion Criteria

We purchased the 2014 Nationwide Readmissions Database (NRD) from Agency for Healthcare Research and Quality (AHRQ) [137]. The NRD is a discharge-level database available for each calendar year from 2010 to 2016 as of November 2019. We chose the 2014 NRD because it was the latest database including comorbidity information. The 2015 and 2016 NRDs do not contain this information. The 2014 NRD is about 12 GB with about 15 million discharges. It was derived from inpatient hospitalization data of 2,048 hospitals in 22 states, which accounted for 51.2% of the U.S. population and 49.3% of all hospitalizations [137]. Patients can be tracked across different hospitals within a state

by the verified and de-identified patient linkage numbers [137]. The 2014 NRD does not have the information of lab tests, vital signs, or medication records.

We constructed four separate cohorts for AMI, COPD, HF, and PN patients with the following criteria: (1) We captured index admissions with a discharge month from January to November in 2014. Index discharges in December were excluded because December was the last month available in the data and 30-day unplanned readmissions cannot be tracked. Eligible index admissions had a principal discharge diagnosis of AMI, COPD, HF, or PN based on the ICD-9-CM codes used by the CMS [5]. The length of stay was longer than 1 day. (2) Patients were older than 18 at admission. They were not transferred to other hospitals and were alive at discharge. They did not leave against medical advice. (3) The gap between index admission discharge and readmission was between one and 30 days (inclusive). Patients could be readmitted for all causes except elective services or other scheduled care according to version 4.0 (ICD-9-CM) of the planned readmission identification criteria used by the CMS [5].

4.2.3 Data Preprocessing and Transformation

The 2014 NRD contains information about admissions, patient demographics, hospital characteristics, diagnoses, and procedures. Diagnoses are available in ICD-9-CM codes and CCS categories [117]. We only kept the CCS-level diagnoses since ICD codes were too granular for this data mining analysis. They are originally represented as one principal diagnosis (DX1) and up to 29 secondary diagnoses (DX2 to DX30). We reshaped diagnoses into a sparse matrix by using the CCS-level diagnoses as attributes and values of true (T) and false (F) representing the presence and the absence of the corresponding

diagnoses. We captured pairs of index admissions of the same patients. For each pair of index admissions, we took the element-wise differences of the readmission status and CCS-level diagnoses. We also extracted patients' characteristics from the first index admission of each pair, including age, sex, primary insurance payer, type of the patient's residing county, and median household income in the patient's zip code. We only used one pair of index admissions of each patient. Attributes of the derived dataset are shown in Table 4.1.

Table 4.1 Attributes of the derived dataset.

Types	Attributes
Fixed factors	Age at the first index admission
	Sex
	Primary insurance payer
	Residing county type
	Median household income in the patient's zip code
Changes	Changes of CCS-level diagnoses' existence
	Change of readmission status

4.2.4 Association Rule Mining

We performed association rule mining [104] to unearth associations rules between changes of diagnoses and the change of readmission status. Association rule mining is an unsupervised data mining approach to discover associations between item sets in the form of "IF {antecedent} THEN {consequent}". We used the Apriori algorithm [138] to identify association rules with the consequent being the change of readmission status (e.g., "readmission → no readmission"). To ensure the association rules were interesting and non-trivial, we used support, confidence, and lift to filter out trivial association rules. The support is the frequency of a pattern occurring in all transactions ranging from 0 to 1. The confidence measures the percentage of transactions with the consequent given it contains

the antecedent. The lift is the ratio of the observed support to expected support given the antecedent and the consequent are independent. A lift greater than 1 indicates the antecedent and the consequent are dependent on each other. Because the standard association rule mining may generate numerous spurious rules [139], we performed Fisher's exact test [140] to measure the statistical significance of the positive correlation between the antecedent and the consequent. We used a significance level of 0.05 and a p-value greater than 0.05 indicated a spurious association rule occurring by chance. For all association rule mining experiments, we used 0.001, 0.75, and 1 as the minimum support, minimum confidence, and minimum lift, respectively. We also assessed each rule's medical soundness and only kept rules with positive associations between a diagnosis change of existence (e.g., "presence (T) → absence (F)" or "absence (F) → presence (T)") and "readmission → no readmission" or "no readmission → readmission".

4.3 Results

4.3.1 Patient Characteristics

We identified 853, 10820, 14343, and 11275 pairs of index admissions from the AMI, COPD, HF, and PN cohorts, respectively. Each pair of index admissions belongs to a unique patient. Table 4.2 shows their demographic information. It can be seen that the majority of patients are older than 65 and are Medicare beneficiaries. There are slightly more female patients. Most patients live in large central metropolitans and neighborhoods with a median household income (zip code level) less than $40,000.

Table 4.2 Demographics of patients in the four cohorts (AMI, COPD, HF, and PN).

Factor	Percentage in Each Cohort			
	AMI	**COPD**	**HF**	**PN**
Age				
18-44	1.8%	2.1%	3.6%	6.7%
45-64	19.9%	34.5%	22.1%	23.5%
65 or older	**78.3%**	**63.4%**	**74.4%**	**69.8%**
Sex				
Female	**52.1%**	**57.8%**	**50.0%**	**51.3%**
Male	47.9%	42.2%	50.0%	48.7%
Primary insurance				
Medicare	**81.8%**	**74.1%**	**78.0%**	**77.8%**
Medicaid	8.0%	13.4%	9.8%	9.8%
Private insurance	7.2%	8.2%	8.2%	9.4%
Other	3.0%	4.2%	4.0%	3.0%
Median household income by zip code				
$1 - $39,999	**32.1%**	**35.2%**	**32.6%**	**29.0%**
$40,000 - $50,999	30.4%	28.9%	26.5%	27.3%
$51,000 - $65,999	19.8%	20.5%	22.1%	23.3%
$66,000 or more	17.7%	15.4%	18.7%	20.3%
Residing county type				
Large central metropolitan	**27.8%**	**26.5%**	**30.9%**	**26.2%**
Large fringe metropolitan	25.7%	24.8%	26.4%	24.6%
Medium metropolitan	21.0%	22.6%	20.5%	22.5%
Small metropolitan	10.0%	10.5%	9.5%	10.7%
Micropolitan	8.8%	8.7%	6.9%	8.9%
Not metropolitan or micropolitan	6.8%	6.8%	5.8%	7.2%

4.3.2 Association Rule Mining

We obtained 108, 72, 81, and 58 eligible association rules for AMI, COPD, HF, and PN cohorts, respectively after applying the rule quality criteria. From each rule, we extracted the diagnosis with changed status (i.e., "T → F" or "F → T") in the antecedent and quantified the strength of its association with readmission by odds ratio. We found 29 diagnoses with a significantly positive association with readmission (risk factor, OR > 1, $P < 0.05$) and showed them in Table 4.3, where "Y" indicates the diagnosis is a risk factor of the cohort. It can be seen from Table 4.3 that only 5/29 (17%) of risk factors are shared by two cohorts and no risk factor is commonly found in three cohorts. This indicates that

the readmission patterns or risk factors of AMI, COPD, HF, and PN cohorts are different. In Appendix 5 – Appendix 8, we showed readmission rates of the four cohorts with and without each risk factor, odd ratio, and 95% confidence interval of OR. For each risk factor that can be potentially modified, we showed one association rule with the highest support in Table 4.4 – Table 4.7.

Table 4.3 Risk factors of readmission of the four cohorts.

Category	Description (CCS-level)	Cohort			
		AMI	**COPD**	**HF**	**PN**
Medical conditions	Chronic ulcer of skin	Y			
	Acute and unspecified renal failure	Y			Y
	Diabetes mellitus without complication	Y	Y		
	Fluid and electrolyte disorders			Y	
	Hemorrhoids				Y
	Hyperplasia of prostate	Y			
	Hypertension with complications and secondary hypertension	Y			
	Infective arthritis and osteomyelitis (except that caused by tuberculosis or sexually transmitted disease)			Y	
	Intestinal obstruction without hernia		Y		
	Late effects of cerebrovascular disease		Y		
	Other connective tissue disease	Y			
	Other diseases of bladder and urethra			Y	
	Other disorders of stomach and duodenum	Y	Y		
	Other gastrointestinal disorders				Y
	Other hereditary and degenerative nervous system conditions		Y		
	Other nutritional; endocrine; and metabolic disorders		Y		
	Pancreatic disorders (not diabetes)			Y	
	Pneumonia (except that caused by tuberculosis or sexually transmitted disease)	Y			
	Retinal detachments; defects; vascular occlusion; and retinopathy	Y			Y
	Rheumatoid arthritis and related disease			Y	
	Septicemia (except in labor)	Y			
	Skin and subcutaneous tissue infections		Y		Y
Mental conditions/ substance-related disorders	Schizophrenia and other psychotic disorders			Y	
	Substance-related disorders	Y			
	Adjustment disorders			Y	
Medical care complications/ adverse effects	Complication of device; implant or graft	Y			
	Complications of surgical procedures or medical care	Y			
	Adverse effects of medical care	Y			
External causes	Superficial injury; contusion				Y

Table 4.4 Association rules of the AMI cohort.

Association Rules	Support	Confidence	Lift	P-value
IF {Hypertension with complications and secondary hypertension = T → F; Sex = Female; Age = 65+} THEN {Readmission → No readmission}	0.041 (35/853)	0.897	1.162	0.036
IF {Other connective tissue disease = T → F; Age = 65+} THEN {Readmission → No readmission}	0.035 (30/853)	0.909	1.177	0.037
IF {Pneumonia (except that caused by tuberculosis or sexually transmitted disease) = T → F; Income = $1-$39,999} THEN {Readmission → No readmission}	0.026 (22/853)	0.957	1.238	0.019
IF {Diabetes mellitus without complication = T → F; Income = $66,000+} THEN {Readmission → No readmission}	0.020 (17/853)	1.000	1.294	0.012
IF {Acute and unspecified renal failure = T → F; Chronic kidney disease = T → T} THEN {Readmission → No readmission}	0.014 (12/853)	1.000	1.294	0.044
IF {Hyperplasia of prostate = F → T; Disorders of lipid metabolism = T → T; Hypertension with complications and secondary hypertension = T → T} THEN {No readmission → Readmission}	0.005 (4/853)	1.000	4.397	0.011
IF {Chronic ulcer of skin = F → T; Congestive heart failure; non-hypertensive = T → T; Location = large central metropolitan} THEN {No readmission → Readmission}	0.005 (4/853)	0.800	3.518	0.011
IF {Substance-related disorders = F → T; Coronary atherosclerosis and other heart disease = T → T; Congestive heart failure; non-hypertensive = T → T} THEN {No readmission → Readmission}	0.004 (3/853)	0.750	3.298	0.011
IF {Septicemia (except in labor) = F → T; Insurance = Medicare} THEN {No readmission → Readmission}	0.004 (3/853)	1.000	4.397	0.003
IF {Other disorders of stomach and duodenum = F → T; Sex = Female; Insurance = Medicare} THEN {No readmission → Readmission}	0.004 (3/853)	0.750	3.298	0.011
IF {Retinal detachments; defects; vascular occlusion; and retinopathy = F → T; Chronic kidney disease = T → T; Sex = Female} THEN {No readmission → Readmission}	0.004 (3/853)	0.750	3.298	0.003

Table 4.5 Association rules of the COPD cohort.

Association Rules	Support	Confidence	Lift	P-value
IF {Other disorders of stomach and duodenum = T → F; Sex = Male} THEN {Readmission → No readmission}	0.002 (23/10820)	0.958	1.225	0.021
IF {Other nutritional; endocrine; and metabolic disorders = T → F; Chronic obstructive pulmonary disease and bronchiectasis = T → T; Cancer of bronchus; lung = T → T; Income = $1-$39,999} THEN {Readmission → No readmission}	0.002 (22/10820)	0.957	1.223	0.026
IF {Late effects of cerebrovascular disease = T → F; Chronic obstructive pulmonary disease and bronchiectasis = T → T; Location = medium metropolitan} THEN {Readmission → No readmission}	0.002 (18/10820)	1.000	1.279	0.012
IF {Intestinal obstruction without hernia = T → F; Coronary atherosclerosis and other heart disease = T → T; Sex = Male} THEN {Readmission → No readmission}	0.001 (13/10820)	1.000	1.279	0.041
IF {Skin and subcutaneous tissue infections = T → F; Screening and history of mental health and substance abuse codes = T → T; Congestive heart failure; non-hypertensive = T → T} THEN {Readmission → No readmission}	0.001 (13/10820)	1.000	1.279	0.041
IF {Other hereditary and degenerative nervous system conditions = T → F; Respiratory failure; insufficiency; arrest (adult) = T → T; Income = $1-$39,999; Sex = Female; Age = 65+} THEN {Readmission → No readmission}	0.001 (13/10820)	1.000	1.279	0.041
IF {Diabetes mellitus without complication = F → T; Other aftercare = T → T; Other nervous system disorders = T → T; Chronic obstructive pulmonary disease and bronchiectasis = T → T} THEN {No readmission → Readmission}	0.001 (11/10820)	0.786	3.607	< 0.001

Table 4.6 Association rules of the HF cohort.

Association Rules	Support	Confidence	Lift	P-value
IF {Schizophrenia and other psychotic disorders = T → F} THEN {Readmission → No readmission}	0.005 (77/14343)	0.865	1.097	0.045
IF {Fluid and electrolyte disorders = T → F; Cancer of prostate = T → T} THEN {Readmission → No readmission}	0.003 (41/14343)	0.976	1.238	0.001
IF {Infective arthritis and osteomyelitis (except that caused by tuberculosis or sexually transmitted disease) = T → F} THEN {Readmission → No readmission}	0.002 (27/14343)	0.931	1.181	0.038
IF {Rheumatoid arthritis and related disease = T → F; Coronary atherosclerosis and other heart disease = T → T; Sex = Female} THEN {Readmission → No readmission}	0.001 (19/14343)	1.000	1.268	0.011
IF {Pancreatic disorders (not diabetes) = T → F; Disorders of lipid metabolism = T → T} THEN {Readmission → No readmission}	0.001 (18/14343)	1.000	1.268	0.014
IF {Other diseases of bladder and urethra = T → F; Disorders of lipid metabolism = T → T; Coronary atherosclerosis and other heart disease = T → T; Chronic kidney disease = T → T; Insurance = Medicare} THEN {Readmission → No readmission}	0.001 (15/14343)	1.000	1.268	0.028
IF {Adjustment disorders = T → F; Location = large central metropolitan; Insurance = Medicare} THEN {Readmission → No readmission}	0.001 (15/14343)	1.000	1.268	0.028

Table 4.7 Association rules of the PN cohort.

Association Rules	Support	Confidence	Lift	P-value
IF {Skin and subcutaneous tissue infections = T → F; Pneumonia (except that caused by tuberculosis or sexually transmitted disease) = T → T; Age = 65+; Insurance = Medicare} THEN {Readmission → No readmission}	0.005 (57/11275)	0.905	1.114	0.035
IF {Retinal detachments; defects; vascular occlusion; and retinopathy = T → F; Sex = Male} THEN {Readmission → No readmission}	0.005 (51/11275)	0.911	1.121	0.035
IF {Acute and unspecified renal failure = T → F; Cancer of prostate = T → T; Insurance = Medicare} THEN {Readmission → No readmission}	0.003 (30/11275)	0.938	1.154	0.045
IF {Other gastrointestinal disorders = T → F; Pneumonia (except that caused by tuberculosis or sexually transmitted disease) = T → T; Cancer of breast = T → T} THEN {Readmission → No readmission}	0.002 (28/11275)	0.966	1.189	0.018
IF {Hemorrhoids = T → F; Essential hypertension = T → T; Location = large fringe metropolitan} THEN {Readmission → No readmission}	0.001 (15/11275)	1.000	1.231	0.044

4.4 Discussions

4.4.1 Analysis of Changes

To our knowledge, this is the first study to analyze changes of modifiable risk factors of readmission with a combination of data mining and statistical methods. Our method has two advantages. First, by comparing the different causes (in ICD code) of index admissions of the same patients, we can better understand readmission through associations between potentially modifiable risk factors and readmission. Second, our method can provide more information about the potential effect of risk factor modification. The traditional logistic regression-based risk factor identification method only evaluates the association between the presence of a factor and the response (e.g., readmission). However, this does not necessarily mean the modification of this risk factor is associated with the reduction of readmission. Our method moves a step further to directly test the association between the changes of risk factors and the change of readmission status.

4.4.2 Potentially Modifiable Risk Factors

By comparing the presence and the absence of the same diagnosis in two different index admissions of the same patient, we can computationally test if the change of the diagnosis is positively associated with the change of readmission status. It is noteworthy that the absence of a diagnosis in the index admission does not necessarily mean the patient is free of the condition, since it is possible that the condition is inactive and does not impact inpatient care [141]. For example, although diabetes mellitus is a nearly incurable chronic condition, this disease can be well-controlled by medications and lifestyle adjustments and changes.

There are four main categories of these 29 risk factors, including medical conditions, mental conditions or substance-related disorders, medical care complications or adverse events, and external causes. We do not consider medical care complications or adverse events as potentially modifiable risk factors because they will only occur after medical care as an outcome. Their risk should be anticipated and minimized during the care process. Similarly, external causes are not patient factors and are not under the control of hospitals. The remaining 25 medical and mental risk factors can be potentially modified by interventions, such as medication, surgery, and psychotherapy. According to the Charlson comorbidity index [85], five of these 25 risk factors have a mortality risk score greater than zero, including acute and unspecified renal failure, chronic ulcer of skin, diabetes mellitus without complication, late effects of cerebrovascular disease, and other connective tissue disease. These conditions are more severe and should receive more attention in the development of interventions.

4.4.3 Recommendation by Association Rules

Table 4.4 – Table 4.7 show that the antecedent of each association rule is composed of patients' non-modifiable factors (e.g., age, sex) and the change of a potentially modifiable risk factor. Because each association rule represents a patient subgroup, it can be used to recommend the modification of readmission risk factors for patients falling into the subgroup. For example, the association rule of "IF {Hypertension with complications and secondary hypertension = T → F; Sex = Female; Age = 65+} THEN {Readmission → No readmission}" in the AMI cohort can be recommended for female AMI patients older than 65 having the comorbidity of hypertension with complications and secondary hypertension. Besides the treatment of the principal condition of AMI, the intervention of

hypertension should be prioritized for these patients to minimize the readmission risk. For each association rule, we performed Fisher's exact test to ensure the association between the antecedent and the consequent is significant. Here, we provide potentially modifiable risk factors instead of intervention plans (e.g., medications, treatment pathways) because we believe that they should be developed by clinicians based on their medical judgments.

4.5 Limitations

Our work has three potential limitations. First, because this is a retrospective analysis, there is no way to control the confounding effects. Although the association rules are represented in "IF-THEN" patterns, the relationship indicates an association not causality. This study cannot replace controlled experiments, such as case-control and prospective cohort studies. However, our results can potentially offer a data-driven hypothesis to a randomized controlled trial and guide other studies to truly disclose the causal relationships between the identified potentially modifiable risk factors and readmission. Second, we used the CCS-level diagnosis in the analysis because ICD-9-CM codes were too granular for data mining analysis. As a result, some diagnoses can only provide very general information, such as "other connective tissue disease". Nevertheless, the purpose of this work is to provide information about potentially modifiable risk factors. Clinicians can potentially map patients' problems and needs into these risk factors and design specific interventions. Third, we relied on ICD codes for identifying the changes of potentially modifiable risk factors. Because ICD codes are mainly designed for billing purposes instead of research, it is possible that some chronic diseases were not recorded in medical records if they did not affect the primary condition. However, this problem would not significantly impact on our findings since our data, the 2014 NRD (derived from

inpatient hospitalization data of 2,048 hospitals in 22 states) is large enough to considerably offset the coding issue since we measured the statistical significance of our findings.

4.6 Conclusions

We performed the analysis of the associations between the changes of potentially modifiable risk factors and the change of readmission status for AMI, COPD, HF, and PN cohorts. To our best knowledge, this finding has not been reported. We identified patterns with potentially modifiable risk factors, and our approach consists of a hybrid of data mining and statistical methods. Compared to existing studies that only consider the impacts of risk factors, our study moves a step further in analyzing the association between potential risk factor modification and readmission prevention. Because each association rule represents a patient subgroup, clinicians can use it to customize interventions for patients falling in the subgroup. In addition, from the association rules, we identified 25 potentially modifiable risk factors of readmission. These modifiable risk factors can be used as potential targets for clinicians to prioritize interventions. Our results would facilitate clinical research to further understand the causes of readmission.

Chapter 5 An Analysis of Orthopedic Patient Satisfaction Survey with Statistical and Data Mining Methods

In this chapter, we retrospectively analyzed results of an orthopedic patient satisfaction survey with statistical and data mining methods. Especially, we applied the analysis method developed in Chapter 4 to study the change of orthopedic patient satisfaction and identified a novel patient-provider sex concordance pattern associated with the change of orthopedic patient satisfaction.

5.1 Background

There were 63 million visits to non-federally employed, office-based orthopedic clinics in the United States in 2010 [142]. Providers are seeking solutions to improve the quality of orthopedic services. Starting in 2012, the CMS in the United States began to adopt patients' perceptions of their hospital experiences in the measurement of care quality [143]. Despite the continued growth of the demand for high-quality orthopedic services, relatively little is known about factors associated with orthopedic patient satisfaction. Only a few studies have explored associations between patient factors and orthopedic patient satisfaction.

Abtahi et al. analyzed surveys of 7,258 patients and found that non-modifiable patient factors, such as age and location can impact orthopedic outpatient satisfaction [144]. Patterson et al. studied surveys of 182 patients and found that patient age and time spent with the provider were associated with increased orthopedic outpatient satisfaction [145].

Menendez et al. found from 150 patients that Spanish language and younger age were predictors of dissatisfaction in an outpatient hand surgery office [146]. Tyser et al. reported that pain, anxiety, and physical function were associated with the patient satisfaction in hand and upper-extremity (non-shoulder) clinic visits based on 1,160 patients [147]. Tisano et al. concluded that Medicare beneficiaries and existing patients were more likely to be satisfied and depression was negatively associated with satisfaction according to surveys of 2,527 patients [148]. Bible et al. reported that younger age, less education, smoking, male patients, and worker's compensation status were associated with dissatisfaction from 200 patients [149]. Hopkins et al. found that high Charlson comorbidity index, increasing elapsed time since surgery or discharge, and increasing length of stay were negatively associated with satisfaction from 1,936 patients after spine surgery [150].

In this chapter, we were interested in identifying patient factors and provider factors associated with orthopedic patient satisfaction. In addition, we wanted to discover factors associated with the change of orthopedic patient satisfaction. Understanding their relationship enables us to design a cost-efficient care quality improvement program, such as customized training for underperforming providers. To accomplish these objectives, we performed a retrospective analysis by multivariate logistic regression, decision tree, and association rule mining.

5.2 Materials and Methods

5.2.1 Ethics

This study was reviewed by the University of Missouri Internal Review Board and qualified for an exemption.

5.2.2 Data Source

This study was a retrospective analysis of the de-identified patient satisfaction survey data collected by the Missouri Orthopedic Institute (MOI), University of Missouri Health Care from 10,136 outpatients after their visits between Jan 11, 2017, and Sep 9, 2018 [151]. The data includes patients' demographic information (sex, age, marital status, city and state of residence, and distance traveled by the patient to MOI), clinical information (height, weight, body mass index (BMI), and diagnosis codes), provider information (de-identified provider number, and provider sex), and visit information (de-identified visit number, de-identified medical record number, visit type, appointment time, survey returned time, survey response time, and answers to nine closed-ended survey questions including the overall satisfaction rating (0-10)).

5.2.3 Data Preprocessing

We removed records with missing values and discretized numeric attributes, including age, distance traveled by the patient to MOI, survey response time, and BMI. The original satisfaction rating was in the 0-10 scale. According to the CMS [152], the 0-10 rating can be classified into "poor" (0, 1, 2), "fair" (3, 4), "good" (5, 6), "very good" (7, 8), and "excellent" (9,10). We further dichotomized the satisfaction rating. Because we were interested in satisfaction improvement, we considered "good" as a neutral response and combined it with "poor" and "fair" into "need to improve" satisfaction. "Very good" and "excellent" were grouped into "no need to improve" satisfaction. We created several new attributes, including parsed appointment time and survey returned time, concordance and combination of patient and provider sex, and if the patient had any secondary diagnoses.

The primary diagnosis codes were converted into CCS categories [117] because ICD-10-CM codes were too granular. Table 5.1 shows the attributes of the preprocessed data set.

Table 5.1 Attributes of the preprocessed data set.

Categories	**Attributes**	
Patient/ provider	Visit type	Survey return day of the month
	Sex	Survey return week of the year
	Race	Survey return month
	Marital status	Survey return year
	Age at discharge (categorized)	Survey response time (categorized)
	City of residence	BMI (categorized)
	State of residence	Diagnosis code 1 (CCS single level)
	Distance to MOI (categorized)	Diagnosis code 1 (CCS level 1)
	Appointment weekday	Diagnosis code 1 (CCS level 2)
	Appointment day of the month	Have secondary diagnoses
	Appointment week of the year	De-identified provider number
	Appointment month	Provider sex
	Appointment year	Patient and provider have the same sex
	Survey return weekday	Patient and provider sex combination
Survey questions	1. Provider explained things in a way that was easy to understand	
	2. Provider gave easy to understand information about health questions or concerns	
	3. Provider knew the important information about patient's medical history	
	4. Provider listened carefully to patient (respondent)	
	5. Provider showed respect for what patient (respondent) had to say	
	6. Provider spent enough time	
	8. Would recommend office	
	9. Rating of provider (binary)	

To analyze the change of satisfaction, we identified pairs of visits of the same patients with different satisfaction ratings. We created a response variable to show the direction of satisfaction change, including "no need to improve satisfaction → need to improve satisfaction", and "need to improve satisfaction → no need to improve satisfaction". For each pair of records (of the same patient), we calculated their difference in numerical variables, such as the difference in age at discharge, the duration between the two appointment dates, and the duration between the two survey return dates. For categorical variables, we recorded their differences between the two visits. We also

included patients' race, sex, age, and marital status at their latest visits in the data. Table 5.2 shows the attributes of the derived satisfaction change data set.

Table 5.2 Attributes of the satisfaction change dataset.

Types	Attributes
Fixed attributes	Age at the latest visit Sex Race Marital status at the latest visit
Numerical difference (discretized)	Difference in age Difference in appointment day Difference in survey return day Difference in survey response time
Categorical difference	Change in patient type Change in primary diagnosis (CCS single level) Change in primary diagnosis (CCS level 1) Change in primary diagnosis (CCS level 2) Change in secondary diagnosis status A different provider (Boolean) Change in provider sex
Response (binary)	Change in satisfaction

5.2.4 Multivariate Logistic Regression Analysis

We used multivariate logistic regression to analyze the associations between independent variables and the binary satisfaction rating. Features were selected by combining forward selection and backward elimination. Candidate models were evaluated by the Akaike information criterion (AIC) [153], the Hosmer-Lemeshow test (HL) [154], and the AUC in 10-fold cross-validation. For all statistical tests, we chose a significance level of 0.05.

5.2.5 Decision Trees

To better understand survey questions' relationship with the overall satisfaction rating, we created a data-driven illustration of the decision path by decision trees. We chose

the C4.5 decision tree algorithm [155] because of its relatively concise tree structure and good performance. The decision trees model's discrimination performance was measured by the AUC in 10-fold cross-validation.

5.2.6 Association Rule Mining

To study associations between patient/provider factor changes and satisfaction changes, we performed association rule mining. We used the Apriori algorithm to identify association rules with the consequent being the change of satisfaction status. In association rule mining, support, confidence, and lift were used to evaluate the strength of rules. Because association rule mining does not necessarily capture statistical dependencies, it may generate spurious association rules by chance. We performed Fisher's exact test to test the significance of the positive correlation between the antecedent and the consequent. For all association rule mining experiments, we set the statistical significance level, minimum support, minimum confidence, and minimum lift to be 0.05, 0.01, 0.9, and 1.0 respectively.

5.3 Results

5.3.1 Patient and Provider Characteristics

After data preprocessing, 10,093 records/visits (8,070 unique patients) were kept in the final data set. The class ratio ("need to improve" satisfaction vs. "no need to improve" satisfaction) was 5.31% to 94.69%. Table 5.3 shows demographic information of the 8,070 patients.

Table 5.3 Demographic information of the 8,070 patients.

Factor	Frequency	Percentage
Age		
0-18	381	4.72%
19-44	1,540	19.08%
45-64	**3,688**	**45.70%**
65+	2,461	30.50%
Sex		
Female	**4,744**	**58.79%**
Male	3,326	41.21%
Race		
African American	519	6.43%
Asian	49	0.61%
Caucasian	**7,337**	**90.92%**
Other	165	2.04%

64 providers were involved in these 10,093 visits. Table 5.4 shows their statistics of service and satisfaction rating by sex. 75.00% (48/64) of them were male providers who accounted for 77.07% (7,779/10,093) of the orthopedic services. The mean numerical ratings (0-10) of female and male providers were not significantly different (Welch's t-test, P = 0.926).

Table 5.4 Providers' statistics of service and satisfaction rating by sex.

Sex	Provider Count	Service Count	Mean Rating	Median Rating	Rating Standard Deviation
Female	16	2,314	9.171	10	1.488
Male	48	7,779	9.167	10	1.552

5.3.2 Relationship between Patient/Provider Factors and Satisfaction

We performed a multivariate logistic regression with patient and provider factors. The response variable was the binary satisfaction rating. After feature selection, nine

independent variables were retained in the final model. The model achieved AIC of 4057.3, AUC of 0.651 (10-fold cross-validation), and calibration of 9.1693 (P = 0.328) where a P-value greater than 0.05 indicates a good fit. Table 5.5 shows the result of this multivariate analysis, including regression coefficient, P-value, odds ratio, and 95% confidence interval of odds ratio. Statistically significant factors (P < 0.05) are bolded. It can be seen that new patients (OR 1.415, 95% CI 1.172-1.708), age between 19 and 44 (OR 3.247, 95% CI 1.573-6.700), age between 45 and 64 (OR 2.239, 95% CI 1.078-4.651), survey returned on Sunday (OR 1.546, 95% CI 1.124-2.128), and having secondary diagnoses (OR 1.226, 95% CI 1.003-1.499) were positively associated with the "need to improve" satisfaction. No negative association was identified.

Table 5.5 Multivariate analysis result (patient and provider factors).

	Coefficient	P-value	OR	95% CI
Patient type				
Existing patient	-	-	-	-
New patient	**0.347**	**<0.001**	**1.415**	**1.172 - 1.708**
Patient sex				
Female	-	-	-	-
Male	-0.186	0.053	0.830	0.688 - 1.002
Patient marital status				
Other marital status	-	-	-	-
Married	-0.190	0.132	0.827	0.646 - 1.059
Single	0.089	0.533	1.094	0.825 - 1.449
Patient age				
0-18	-	-	-	-
19-44	**1.178**	**0.001**	**3.247**	**1.573 - 6.700**
45-64	**0.806**	**0.031**	**2.239**	**1.078 - 4.651**
65+	0.152	0.695	1.164	0.546 - 2.483
City				
Other cities	-	-	-	-
Columbia, MO	0.503	0.131	1.654	0.861 - 3.178
Distance to MOI				
0-20 miles	-	-	-	-
20.1-33 miles	0.618	0.070	1.854	0.951 - 3.615
33.1-74 miles	0.643	0.054	1.903	0.988 - 3.665
74.1+ miles	0.444	0.189	1.559	0.803 - 3.024
Provider				
Provider #1	-	-	-	-
Provider #2	0.063	0.953	1.065	0.131 - 8.641
Provider #3	-0.494	0.646	0.610	0.074 - 5.018
…	…	…	…	…
Survey return day				
Monday	-	-	-	-
Tuesday	0.263	0.123	1.301	0.931 - 1.818
Wednesday	0.229	0.168	1.258	0.908 - 1.743
Thursday	-0.006	0.972	0.994	0.706 - 1.400
Friday	0.001	0.997	1.001	0.718 - 1.394
Saturday	0.087	0.610	1.091	0.781 - 1.524
Sunday	**0.436**	**0.007**	**1.546**	**1.124 - 2.128**
Secondary diagnoses				
No	-	-	-	-
Yes	**0.204**	**0.047**	**1.226**	**1.003 - 1.499**

5.3.3 Relationship between Survey Questions and Satisfaction

The C4.5 decision tree model had good discrimination performance (AUC, 0.891) in 10-fold cross-validation. The resulting decision tree is shown in Figure 5.1. We abbreviated the names of questions and answers. The circles represent questions. For

example, “Q8” means question 8. Table 5.1 shows all the survey questions. The texts on the decision paths are the answers to the corresponding questions. The possible answers are “yes, definitely” (“YD”), “yes, somewhat” (“YS”), and “no” (“N”). The rectangles represent leaf nodes (class labels or the values of the dependent variable). In this analysis, we used the binary response variable with “B” indicating the “need to improve” satisfaction and “G” meaning there is “no need to improve” satisfaction. The first number in the parenthesis is the number of records reaching the leaf and the second number indicates the misclassified cases. For example, in the decision path of “if Q8 = N then the class is “B” (308/30)”, 308 records reached this leaf and 30 of them were misclassified.

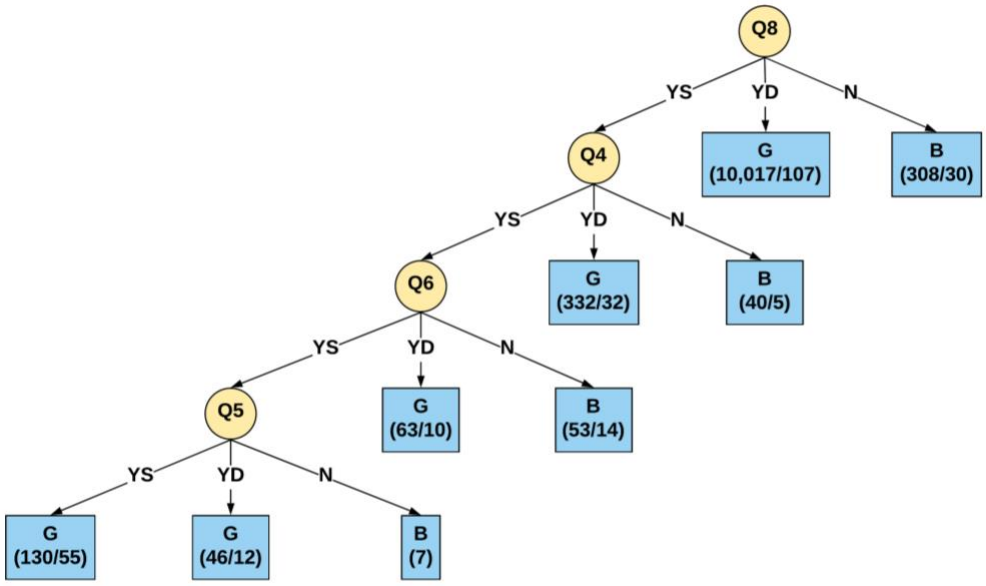


Figure 5.1 Decision tree for the impact of survey questions on the overall provider rating.

We also performed a multivariate logistic regression analysis on the same data. After feature selection, questions 1, 3-6, and 8 were kept in the final model. The model

achieved AIC of 1828.7, AUC of 0.931 (10-fold cross-validation), and calibration (HL chi-squared) of 11.036 (P = 0.200). The result is shown in Table 5.6.

Table 5.6 Multivariate analysis result (survey questions only).

	Coefficient	P-value	OR	95% CI
Intercept	**-4.832**	**<0.001**	**0.008**	**0.006 - 0.010**
Question 1				
Yes, definitely	-	-	-	-
Yes, somewhat	**0.374**	**0.043**	**1.454**	**1.013 - 2.088**
No	**1.835**	**<0.001**	**6.265**	**2.444 - 16.059**
Question 3				
Yes, definitely	-	-	-	-
Yes, somewhat	0.247	0.167	1.280	0.902 - 1.815
No	**0.771**	**0.005**	**2.163**	**1.270 - 3.683**
Question 4				
Yes, definitely	-	-	-	-
Yes, somewhat	**0.697**	**0.001**	**2.007**	**1.333 - 3.023**
No	**1.960**	**<0.001**	**7.099**	**2.904 - 17.356**
Question 5				
Yes, definitely	-	-	-	-
Yes, somewhat	**0.714**	**<0.001**	**2.042**	**1.378 - 3.027**
No	**1.912**	**<0.001**	**6.768**	**2.800 - 16.356**
Question 6				
Yes, definitely	-	-	-	-
Yes, somewhat	**0.755**	**<0.001**	**2.128**	**1.466 - 3.090**
No	**1.845**	**<0.001**	**6.330**	**3.604 - 11.120**
Question 8				
Yes, definitely	-	-	-	-
Yes, somewhat	**1.751**	**<0.001**	**5.760**	**3.977 - 8.342**
No	**3.997**	**<0.001**	**54.410**	**31.569 - 93.776**

5.3.4 Satisfaction Change Analysis

We identified 175 patients who gave different satisfaction feedback for different visits and created 203 records to represent the pairwise difference between the same patient's different visits. The Apriori algorithm generated 240 and 196 significant and non-redundant rules with the consequent of "need to improve satisfaction → no need to improve satisfaction" and "no need to improve satisfaction → need to improve satisfaction", respectively. We found eight interesting rules and displayed them in Table 5.7. The first

four rules have the consequent of “need to improve satisfaction → no need to improve satisfaction”. These four association rules have the factor of “male provider → female provider” in their antecedents. Especially for female patients with other factors unchanged, changing from a male provider to a female provider was associated with the change from “need to improve” satisfaction to “no need to improve” satisfaction. The last four association rules in Table 5.7 have the consequent of “no need to improve satisfaction → need to improve satisfaction”. Three of them indicate that for female patients with other factors unchanged, switching a female provider to a male provider was associated with the change from “no need to improve” satisfaction to “need to improve” satisfaction. This sex concordance pattern also applies to male patients. One rule indicates that for a male patient, changing a male provider to a female provider was associated with the change from “no need to improve” satisfaction to “need to improve” satisfaction. These rules indicated the association between the change of provider sex and the change of satisfaction. By checking the de-identified medical record numbers, we verified that these rules were extracted from different patients. These eight rules have perfect confidence (100%), lift greater than 1.0, and P-value lower than 0.05, indicating significantly positive correlations.

Table 5.7 Eight interesting association rules.

Association Rule	Support	Confidence	Lift	P-value
IF {Single patient, the survey was returned within 1 to 3 months after the last submission, male provider → female provider} THEN {Need to improve satisfaction → No need to improve satisfaction}	0.030 (6/203)	1.000	1.845	0.024
IF {Single patient, the patient revisited MOI within 1 to 3 months after the last appointment, male provider → female provider} THEN {Need to improve satisfaction → No need to improve satisfaction}	0.030 (6/203)	1.000	1.845	0.024
IF {No change in patient type, no change in secondary diagnosis status, the survey was returned within 1 to 3 months after the last submission, female patient, male provider → female provider} THEN {Need to improve satisfaction → No need to improve satisfaction}	0.025 (5/203)	1.000	1.845	0.045
IF {No change in primary diagnosis (CCS level 1), single patient, Caucasian patient, female patient, male provider → female provider} THEN {Need to improve satisfaction → No need to improve satisfaction}	0.025 (5/203)	1.000	1.845	0.045
IF {Patient age was between 45 and 64, Caucasian patient, female patient, female provider → male provider} THEN {No need to improve satisfaction → Need to improve satisfaction}	0.030 (6/203)	1.000	2.183	0.008
IF {No change in patient age, Caucasian patient, no change in patient type, female patient, female provider → male provider} THEN {No need to improve satisfaction → Need to improve satisfaction}	0.025 (5/203)	1.000	2.183	0.019
IF {No change in primary diagnosis (CCS level 1), no change in patient type, Caucasian patient, female patient, female provider → male provider} THEN {No need to improve satisfaction → Need to improve satisfaction}	0.025 (5/203)	1.000	2.183	0.019
IF {No change in patient age, no change in primary diagnosis (CCS single level), male patient, male provider → female provider} THEN {No need to improve satisfaction → Need to improve satisfaction}	0.020 (4/203)	1.000	2.183	0.043

5.4 Discussions

5.4.1 Predictors of Patient Satisfaction

We found that being a new patient (OR 1.415, 95% CI 1.172-1.708) was positively associated with the "need to improve" satisfaction compared to an existing patient. This result was similar to the finding by Tisano et al. [148] that existing patients were more satisfied than new patients. Patients may return to MOI because of positive experience before and this may influence their ratings for current visits. Another interpretation was that question 3 asked if the provider knew the important information about the patient's medical history. New patients will mostly answer "no" to this question and this may impact the provider's overall rating.

Young adults (19-44) (OR 3.247, 95% CI 1.573-6.7) and middle-aged adults (45-64) (OR 2.239, 95% CI 1.078-4.651) were positively associated with the "need to improve" satisfaction. We did not find a statistically significant association between older adults (65+) and the "need to improve" satisfaction. Previous studies [156,157] showed that older patients were more satisfied than younger patients. It has been reported that older patients care less about the provider's attitude and friendliness than younger patients [158], and older patients tend to skip questions rather than give low scores [159].

Additionally, surveys returned on Sunday (OR 1.546, 95% CI 1.124-2.128) were positively associated with the "need to improve" satisfaction when compared with the baseline level "survey returned on Monday". To the extent of our knowledge, no study has reported the relationship between a survey return day and satisfaction. According to a psychological study of the relationship between mood patterns and the day of the week,

Sunday has the lowest positive affect activation, which contains four components of "active", "attentive", "inspired", and "interested" [160]. This may contribute to decreased satisfaction.

Having secondary diagnoses (OR 1.226, 95% CI 1.003-1.499) was more positively associated with the "need to improve" satisfaction than not having secondary diagnoses. Having secondary diagnoses indicates the patient has comorbidities and/or complications. The impact of comorbidities on orthopedic patient satisfaction is still unclear and under debate. Our result agreed with the study by Husted et al. [161], which showed that patients without comorbidities were more satisfied than those with comorbidities following total hip or knee replacement. Bourne et al. [162] reported that postoperative complications were associated with patient dissatisfaction after total knee replacement.

Patients' primary diagnosis in three CCS levels was excluded from the final model by feature selection. This indicated that the primary diagnosis did not contribute to the classification of patient satisfaction. We did not find evidence from the literature that patient satisfaction was related to any specific primary diagnosis. However, some studies reported the linkage between the severity of certain conditions and satisfaction although the results were contradicting. For example, Kavalnienė et al. [163] reported that patients' satisfaction with primary care is negatively associated with the severity of depression. Schmocker et al. [164] found that the severity of disease did not impact satisfaction for patients with diverticulitis.

5.4.2 Relationship between Survey Questions and Satisfaction Rating

We explored the relationship between survey questions and the overall satisfaction rating with the C4.5 decision tree algorithm. It can be seen from Figure 5.1 that question 8 is the most decisive because its answers are the condition of all decision paths. One possible reason is that question 8 asks if the patient would recommend the office and it is nearly equivalent to asking about patient satisfaction (the dependent variable). The decision tree has a pattern that for each question, the answer of "no" (N) always reaches the leaf node of "need to improve" satisfaction (B), and the answer of "yes, definitely" (YD) always points to "no need to improve" satisfaction (G). For the less certain answer of "yes, somewhat" (YS), questions 4, 6, and 8 (Q4, Q6, & Q8) always rely on other questions to finish the decision. Question 5 can reach "no need to improve" satisfaction (G) with the answer of "yes, somewhat" (YS). This finding indicates that it is essential for providers to treat patients with respect during patient encounters for better patient satisfaction. Questions 1-3 were not found in the decision tree meaning these questions were less associated with the decision/satisfaction.

To further understand each question's impact on the overall satisfaction, we performed a multivariate logistic regression analysis. The result (see Table 5.6) indicates that the answers of "yes, somewhat" and "no" to questions 1, 4-6, and 8 were more positively associated with "need to improve" satisfaction than "yes, definitely". For question 3, we only observed this relationship on the answer of "no". Based on the decision tree and the logistic regression analysis, providers' interaction and respect for patients play an important role in the overall satisfaction rating.

5.4.3 Satisfaction Change

To our knowledge, this is the first data mining analysis of orthopedic patient satisfaction change. The most interesting finding was the association between the change of provider sex and the change of satisfaction. For female patients, the association holds in both directions (male provider to female provider → "need to improve" satisfaction to "no need to improve" satisfaction; female provider to male provider → "no need to improve" satisfaction to "need to improve" satisfaction). For male patients, the association holds in one direction only (male provider to female provider → "no need to improve" satisfaction to "need to improve" satisfaction).

The impact of patient and provider sex concordance on patient satisfaction has been studied in other care specialties, but the findings were conflicting. Schmittdiel et al. [165] reported that female patients of female providers were the least satisfied compared to other sex combinations in preventive care. Rogo-Gupta et al. [166] found that female gynecologists were significantly less likely to receive top satisfaction scores than male counterparts. Derose et al. [167] reported that female patients in the emergency department trusted female physicians more.

5.5 Limitation

The satisfaction change analysis has a relatively smaller sample size compared to the other two analyses because the change of satisfaction is not a frequent event. However, association rule mining has no assumptions about sample size and distribution. We also performed Fisher's exact test to ensure our findings were statistically significant.

5.6 Conclusions

In this chapter, we performed a retrospective analysis of the orthopedic patient satisfaction data. We found that new patients, young adults (19 - 44), middle-aged adults (45 - 64), survey returned on Sunday, and having secondary diagnoses were positively associated with the “need to improve” satisfaction. To the best of our knowledge, no prior studies have reported the association between a survey return day and orthopedic patient satisfaction. Providers’ interaction and respect for patients were found to be important factors in patient satisfaction. By association rule mining, we found eight interesting association rules that can help to explain patients' satisfaction changes. The key finding was that the change of provider sex was associated with the satisfaction change. To better understand patient satisfaction, more information about patients and providers is needed, such as patients’ previous satisfaction responses, socioeconomic status and detailed medical history, providers’ specialty and previous experience. Further studies are necessary to disclose the true reasons for satisfaction change.

Chapter 6 Conclusions

6.1 Summary of Findings

The overall objective of this project was to improve the timeliness of unplanned 30-day hospital readmission preventive intervention through a data-driven approach. To achieve this objective, we used various predictive modeling and exploratory analysis methods to develop an early prediction model of readmission and identify potentially modifiable risk factors associated with readmission.

In Chapter 2, we performed a systematic review of readmission risk factors from 13 articles and identified 34 highly generalizable risk factors in seven categories, including social factors, healthcare utilization, index admission characteristics, comorbidities, lab test results, medications, and hospital characteristics. We determined that existing studies tended to ignore detailed medical history and only used high-level information about previous healthcare utilization (e.g., the number of inpatient visits in last year). In addition, to improve readmission predictive models' performance, some studies sacrificed their models' timeliness by including variables whose values would only become available near or after discharge. As a result, most reported predictive models cannot predict readmission risk at the early stage of inpatient care.

In Chapter 3, we developed an early prediction model of readmission by incorporating detailed medical history (diagnoses, procedures, medications, lab test results, vital signs, and healthcare utilization) up to one year before the time point of prediction in the predictive model. We also collected information that can become available in EHR

within 24 hours of the current inpatient admission, including patients' demographic information, lab test results, vital signs, and medications. The model was derived and validated on 96,550 patients' data identified from 205 hospitals in the United States using various statistical and machine learning algorithms. The XGBoost model has the best performance among the six candidate models. Its performance is better than the HOSPITAL score, the LACE index, and the LACE-rt index for all of the five performance metrics on the validation set. In addition, we identified 14 novel risk factors and two novel protective factors of readmission by multivariate analysis. The early prediction model can potentially help clinicians to identify readmission risk at the early stage of hospitalization so that clinicians can have more time to pay extra attention to high-risk patient's care.

In Chapter 4, we used association rule mining and statistical methods to identify potentially modifiable risk factors for readmission of four patient cohorts eligible for the HRRP, including AMI, COPD, HF, and PN. We developed a novel method to compare diagnoses and readmission status of the same patients' different index admissions. This method allowed us to better disclose associations between potentially modifiable risk factors and readmission. This method can also provide more information about the potential effect of risk factor modification. Compared to the traditional logistic regression-based method that can only evaluate the association between the presence of a factor and the response (e.g., readmission), our method moves a step further to directly test the association between changes of risk factors and the change of readmission status. We identified 29 risk factors of readmission in four categories, including medical conditions, mental conditions or substance-related disorders, medical care complications or adverse events, and external causes. 25 of them were potentially modifiable. Because each

association rule containing the potentially modifiable risk factor represents a patient subgroup, clinicians can use it to customize interventions for patients falling in the subgroup.

In Chapter 5, we analyzed results of an orthopedic patient satisfaction survey with statistical and data mining methods. We identified one novel risk factor of the "need to improve" satisfaction and found that providers' interaction and respect for patients were important factors in satisfaction. We applied the analysis method developed in Chapter 4 to study factors associated with the change of orthopedic patient satisfaction and identified a novel patient-provider sex concordance pattern. For both female and male patients, the change of a provider of the same sex into a provider of the different sex was associated with the change from satisfaction to dissatisfaction. This pattern can potentially be used to improve orthopedic patient satisfaction.

6.2 Limitation

This project was a retrospective analysis. The early prediction model was developed based on historical data. Similarly, the identified relationship between changes of potentially modifiable risk factors and the change of readmission status was association, not causality. However, our results can potentially offer a data-driven hypothesis to a randomized controlled trial and guide other studies to truly disclose the causal relationships between the identified risk factors and readmission.

6.3 Contributions

The outcome of this project can potentially shift the timing of readmission preventive intervention to the early stage of hospitalization. Clinicians can potentially use

the early prediction model to stratify patients' readmission risk at the beginning of care and allocate more resources to high-risk patients during the care process. The association rules can potentially direct clinicians to plan interventions for different patients based on their social and clinical factors during hospitalization. Although this project focused on the context of readmission, the methodology can be generalized to other research topics, such as the recurrence of a disease.

APPENDICES

Appendix 1. Implementations of the HOSPITAL score & the LACE/LACE-rt index.

The point system of the HOSPITAL score [28].

HOSPITAL Score	Points
Hemoglobin level at discharge	
≥ 12 g/dL	0
< 12 g/dL	1
Discharge from an oncology service	
No	0
Yes	2
Sodium level at discharge	
≥ 135 mEq/L	0
< 135 mEq/L	1
Procedure (ICD) during hospital stay	
No	0
Yes	1
Index admission type	
Elective	0
Nonelective	1
Number of hospital admissions last year	
0	0
1 – 5	2
> 5	5
Length of stay	
< 5 d	0
≥ 5 d	2

The point system of the LACE/LACE-rt index [27].

LACE/LACE-rt Index	Points
Length of stay (current admission for LACE, previous admission within past 30 days for LACE-rt)	
< 1 d	0
1 d	1
2 d	2
3 d	3
4 – 6 d	4
7 – 13 d	5
≥ 14 d	7
Acute admission	
No	0
Yes	3
Charlson comorbidity index score	
0	0
1	1
2	2
3	3
≥ 4	5
Number of emergency department visits during previous six months	
0	0
1	1
2	2
3	3
≥ 4	4

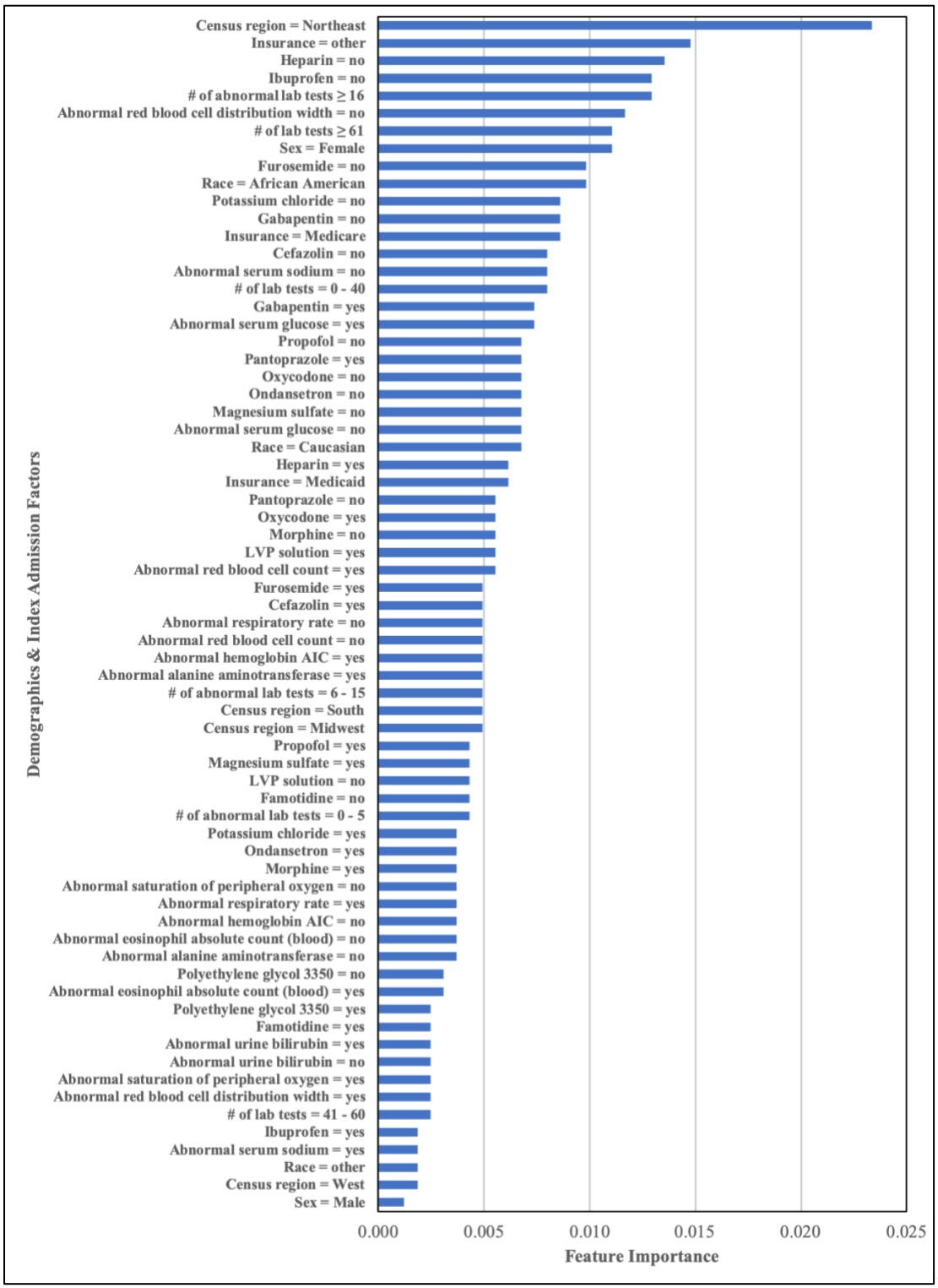
Python implementation of the HOSPITAL score.

```
def HOSPITAL_SCORE(record):

    """
    Python implementation of the HOSPITAL score.

    Variables:
    @ record: A Pandas dataframe row that holds all elements of a record
    @ total_score: The cumulative HOSPITAL score

    """

    total_score = 0

    if record['Hemoglobin_level_at_discharge'] < 12:
        total_score += 1

    if record['Discharge_from_an_oncology_service'] == 'Yes':
        total_score += 2

    if record['Sodium_level_at_discharge'] < 135:
        total_score += 1

    if record['ICD_procedure_during_stay'] == 'Yes':
        total_score += 1

    if record['Index_admission_type'] == 'Nonelective':
        total_score += 1

    if record['N_admissions_last_year'] == 0:
        total_score += 0
    elif record['N_admissions_last_year'] >= 1 and record['N_admissions_last_year'] <= 5:
        total_score += 2
    else:
        total_score += 5

    if record['length_of_stay'] >= 5:
        total_score += 2

    return total_score
```

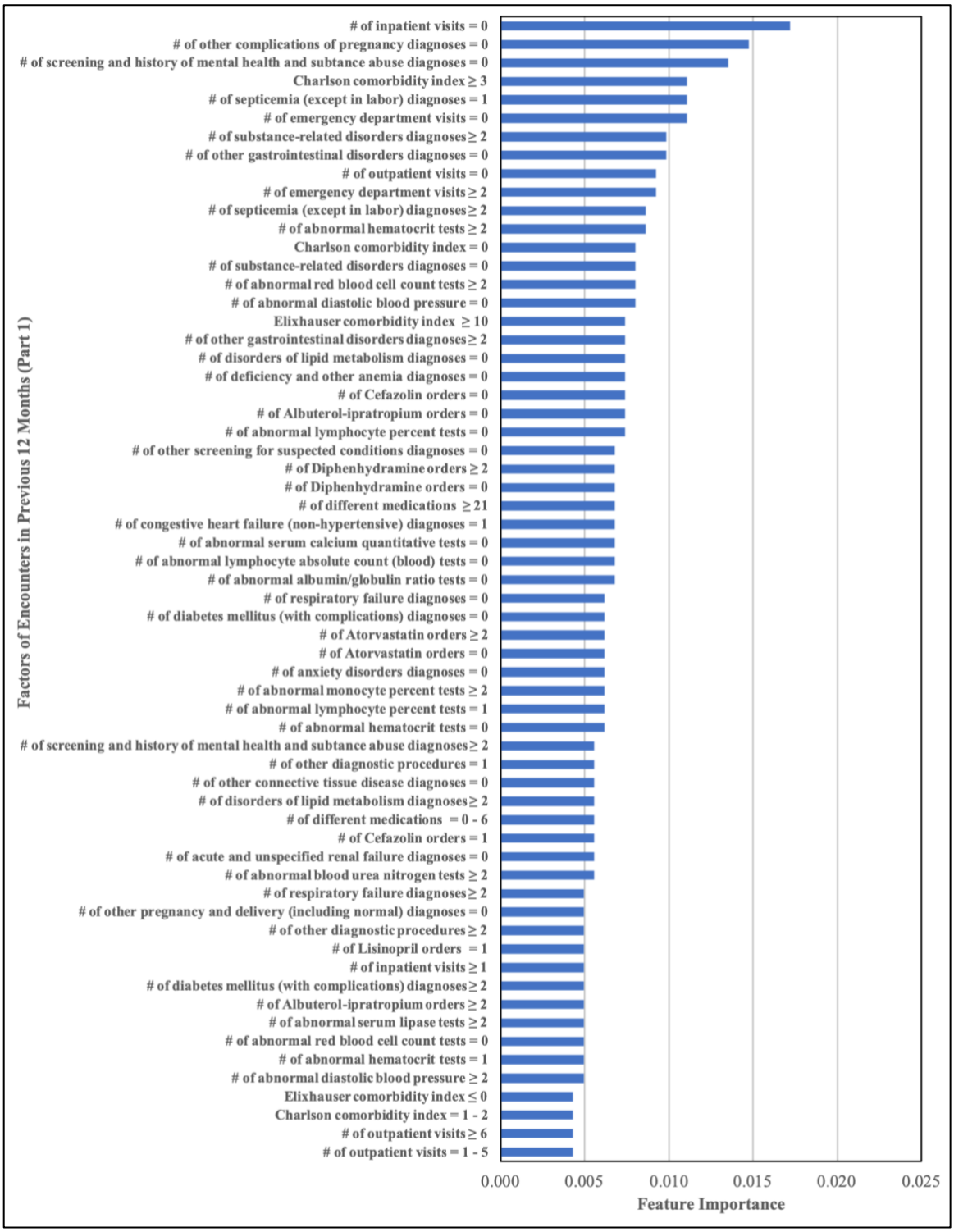
Python implementation of the LACE/LACE-rt index.

```
def LACE_INDEX(record, LACE_rt = False):

    """
    Python implementation of the original LACE index (by default) and the LACE-rt index.

    Variables:
    @ record: A Pandas dataframe row that holds all elements of a record
    @ LACE_rt: The switch between the orginal LACE index and the LACE-rt index
    @ total_score: The cumulative LACE or LACE-rt score
    @ length_of_stay: The type of length of stay
        For the LACE index, it is for the current admission
        For the LACE-rt index, it is for the previous admission within last 30 days

    """

    total_score = 0

    length_of_stay = 'Length_of_stay_current_admission'

    if LACE_rt:
        length_of_stay = 'Length_of_stay_previous_admission_past_30_days'

    if record[length_of_stay] < 1:
        total_score += 0
    elif record[length_of_stay] == 1:
        total_score += 1
    elif record[length_of_stay] == 2:
        total_score += 2
    elif record[length_of_stay] == 3:
        total_score += 3
    elif record[length_of_stay] >= 4 and record[length_of_stay] <= 6:
        total_score += 4
    elif record[length_of_stay] >= 7 and record[length_of_stay] <= 13:
        total_score += 5
    else:
        total_score += 7

    if record['Acute_admission'] == 'Yes':
        total_score += 3

    if record['Charlson_comorbidity_index'] == 0:
        total_score += 0
    elif record['Charlson_comorbidity_index'] == 1:
        total_score += 1
    elif record['Charlson_comorbidity_index'] == 2:
        total_score += 2
    elif record['Charlson_comorbidity_index'] == 3:
        total_score += 3
    else:
        total_score += 5

    if record['N_emergency_department_visits_past_6_months'] == 0:
        total_score += 0
    elif record['N_emergency_department_visits_past_6_months'] == 1:
        total_score += 1
    elif record['N_emergency_department_visits_past_6_months'] == 2:
        total_score += 2
    elif record['N_emergency_department_visits_past_6_months'] == 3:
        total_score += 3
    else:
        total_score += 4

    return total_score
```

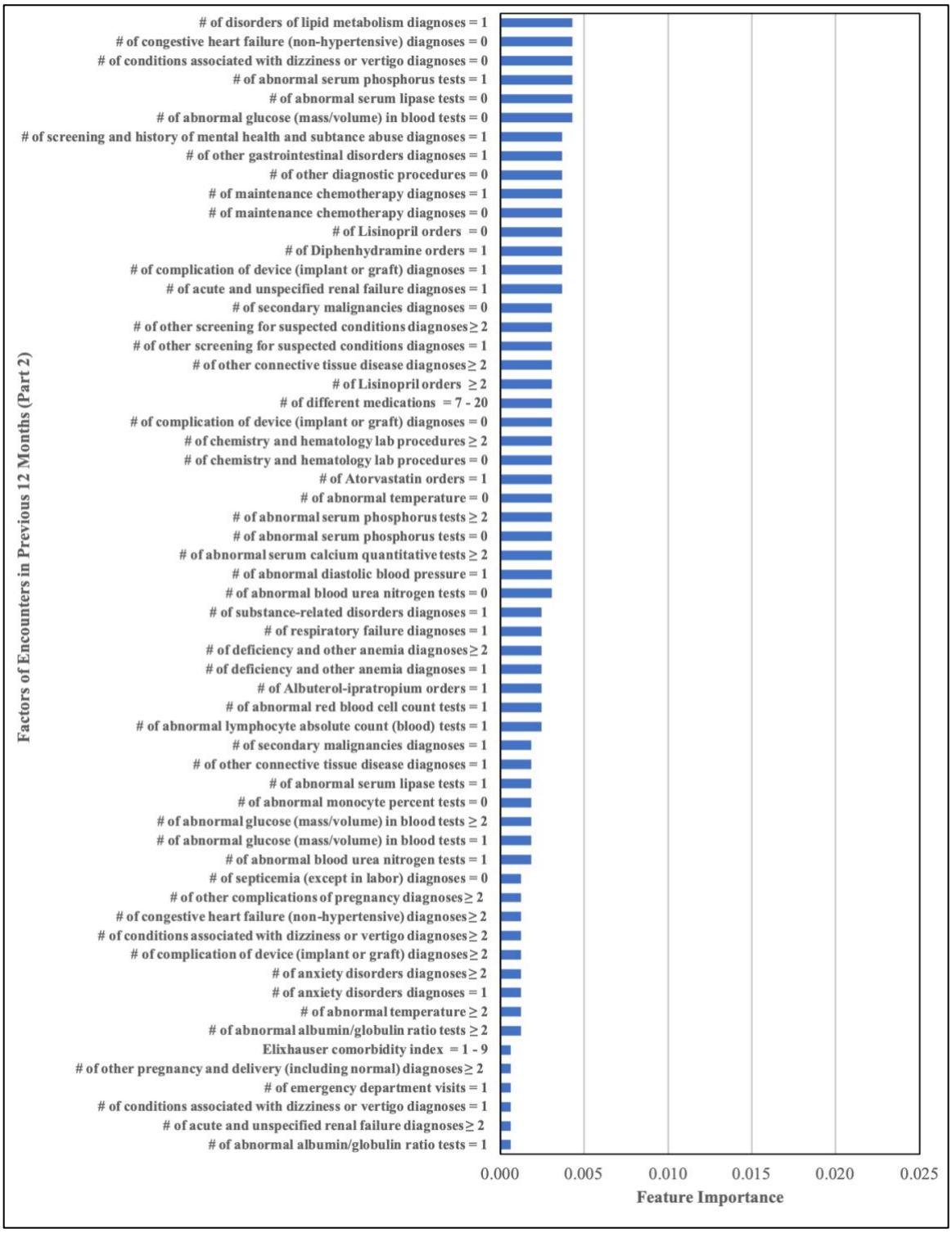
Appendix 2. Demographics and index admission factors of the XGBoost model.



Appendix 3. Medical history (last 12 months) factors of the XGBoost Model (Part 1).



Appendix 4. Medical history (last 12 months) factors of the XGBoost Model (Part 2).



Appendix 5. Risk factors of the AMI cohort.

Diagnosis (Dx)	Readmission Rate		OR	95% CI
	w/ Dx	w/o Dx		
Chronic ulcer of skin	24.5%	13.7%	2.043	1.849 - 2.258
Other disorders of stomach and duodenum	20.5%	14.0%	1.584	1.337 - 1.877
Septicemia (except in labor)	19.4%	14.0%	1.483	1.290 - 1.705
Acute and unspecified renal failure	19.3%	12.7%	1.647	1.570 - 1.728
Pneumonia (except that caused by tuberculosis or sexually transmitted disease)	18.7%	13.6%	1.458	1.363 - 1.559
Hypertension with complications and secondary hypertension	18.5%	11.8%	1.690	1.619 - 1.765
Complications of surgical procedures or medical care	17.2%	14.0%	1.274	1.109 - 1.463
Retinal detachments; defects; vascular occlusion; and retinopathy	17.2%	14.0%	1.281	1.142 - 1.436
Adverse effects of medical care	16.9%	14.0%	1.251	1.113 - 1.407
Hyperplasia of prostate	15.9%	13.9%	1.166	1.074 - 1.265
Other connective tissue disease	15.6%	13.9%	1.149	1.076 - 1.228
Substance-related disorders	15.6%	14.0%	1.138	1.016 - 1.275
Diabetes mellitus without complication	15.4%	13.6%	1.158	1.105 - 1.212
Complication of device; implant or graft	15.1%	14.0%	1.096	1.006 - 1.194

Appendix 6. Risk factors of the COPD cohort.

Diagnosis (Dx)	Readmission Rate		OR	95% CI
	w/ Dx	w/o Dx		
Skin and subcutaneous tissue infections	25.5%	18.5%	1.513	1.416 - 1.616
Intestinal obstruction without hernia	22.7%	18.6%	1.286	1.126 - 1.469
Other disorders of stomach and duodenum	21.6%	18.5%	1.208	1.099 - 1.327
Late effects of cerebrovascular disease	21.6%	18.5%	1.210	1.132 - 1.294
Other hereditary and degenerative nervous system conditions	20.5%	18.5%	1.136	1.079 - 1.196
Other nutritional; endocrine; and metabolic disorders	20.1%	18.0%	1.146	1.124 - 1.169
Diabetes mellitus without complication	19.4%	18.2%	1.079	1.058 - 1.101

Appendix 7. Risk factors of the HF cohort.

Diagnosis (Dx)	Readmission Rate		OR	95% CI
	w/ Dx	w/o Dx		
Schizophrenia and other psychotic disorders	27.1%	19.8%	1.510	1.415 - 1.612
Infective arthritis and osteomyelitis (except that caused by tuberculosis or sexually transmitted disease)	26.4%	19.8%	1.452	1.279 - 1.649
Pancreatic disorders (not diabetes)	23.4%	19.8%	1.236	1.093 - 1.396
Adjustment disorders	23.1%	19.8%	1.214	1.037 - 1.421
Other diseases of bladder and urethra	22.5%	19.8%	1.173	1.074 - 1.282
Fluid and electrolyte disorders	21.8%	18.9%	1.197	1.176 - 1.218
Rheumatoid arthritis and related disease	21.2%	19.8%	1.089	1.026 - 1.156

Appendix 8. Risk factors of the PN cohort.

Diagnosis (Dx)	Readmission Rate		OR	95% CI
	w/ Dx	w/o Dx		
Skin and subcutaneous tissue infections	19.0%	14.4%	1.399	1.324 - 1.479
Hemorrhoids	17.9%	14.4%	1.293	1.165 - 1.435
Other gastrointestinal disorders	17.6%	13.6%	1.355	1.329 - 1.382
Acute and unspecified renal failure	17.4%	13.9%	1.309	1.281 - 1.338
Retinal detachments; defects; vascular occlusion; and retinopathy	16.9%	14.4%	1.206	1.138 - 1.278
Superficial injury; contusion	15.8%	14.4%	1.112	1.021 - 1.212

BIBLIOGRAPHY

1. National Quality Forum. Endorsement Summary: All-Cause Readmissions [Internet]. 2012. Available from: http://www.qualityforum.org/Projects/Readmissions_Endorsement_Maintenance.aspx

2. The Centers for Medicare & Medicaid Services. Hospital Quality Initiative - Outcome Measures 2016 Chartbook [Internet]. 2016. Available from: https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/OutcomeMeasures.html

3. Jencks SF, Williams M V., Coleman EA. Rehospitalizations among Patients in the Medicare Fee-for-Service Program. New England Journal of Medicine 2009;360(14):1418–1428. PMID: 19339721

4. Medicare. Readmissions Reduction Program (HRRP) [Internet]. 2012. Available from: https://www.cms.gov/medicare/medicare-fee-for-service-payment/acuteinpatientpps/readmissions-reduction-program.html

5. Yale New Haven Health Services Corporation/Center for Outcomes Research & Evaluation. 2016 Condition-Specific Measures Updates and Specifications Report Hospital-Level 30-Day Risk-Standardized Readmission Measures. 2016.

6. Hansen LO, Young RS, Hinami K, Leung A, Williams M V. Interventions to Reduce 30-day Rehospitalization: a Systematic Review. Annals of internal medicine 2011 Oct 18;155(8):520–8. PMID: 22007045

7. Medicare. Hospital Compare - Unplanned Hospital Visits [Internet]. 2009. Available from: https://www.medicare.gov/hospitalcompare/About/HospitalReturns.html

8. Medicare. Hospital Inpatient Quality Reporting Program [Internet]. 2003. Available from: https://www.cms.gov/medicare/quality-initiatives-patient-assessment-instruments/hospitalqualityinits/hospitalrhqdapu.html

9. The 111th United States Congress. The Patient Protection and Affordable Care Act [Internet]. 2010. Available from: https://www.govinfo.gov/content/pkg/PLAW-111publ148/pdf/PLAW-111publ148.pdf

10. The Centers for Medicare & Medicaid Services. FAQs: CMS Publically Reported Risk-Standardized Outcome and Payment Measures. 2014.

11. Boccuti C, Casillas G. Aiming for Fewer Hospital U-turns : The Medicare Hospital Readmission Reduction Program. Policy Brief. Menlo Park, CA; 2017.

12. Fontana E, Hawes K. Map: See the 2,599 Hospitals That Will Face Readmissions Penalties This Year [Internet]. The Advisory Board. 2018. Available from:

https://www.advisory.com/daily-briefing/2018/09/27/readmissions

13. Rau J. New Round of Medicare Readmission Penalties Hits 2,583 Hospitals [Internet]. Kaiser Health News. 2019. Available from: https://khn.org/news/hospital-readmission-penalties-medicare-2583-hospitals/

14. Rau J. 2,573 Hospitals Will Face Readmission Penalties This Year. Is Yours One of Them? [Internet]. Kaiser Health News. 2017. Available from: https://www.advisory.com/daily-briefing/2017/08/07/hospital-penalties

15. Kripalani S, Theobald CN, Anctil B, Vasilevskis EE. Reducing Hospital Readmission Rates: Current Strategies and Future Directions. Annual Review of Medicine 2014;65:471–85. PMID: 24160939

16. Ballas SK, Lusardi M. Hospital Readmission for Adult Acute Sickle Cell Painful Episodes: Frequency, Etiology, and Prognostic Significance. American journal of hematology 2005 May;79(1):17–25. PMID: 15849770

17. Schall MB, Flannery J. Undertreatment of Women with Heart Failure: a Reversible Outcome on Hospital Readmission. Lippincott's Case Management : Managing the Process of Patient Care 2004 Nov;9(6):250–3. PMID: 15602332

18. Gustafsson F, Torp-Pedersen C, Burchardt H, Buch P, Seibaek M, Kjøller E, et al. Female Sex Is Associated with a Better Long-term Survival in Patients Hospitalized with Congestive Heart Failure. European Heart Journal 2004 Jan;25(2):129–35. PMID: 14720529

19. Emerson CB, Eyzaguirre LM, Albrecht JS, Comer AC, Harris AD, Furuno JP. Healthcare-associated infection and hospital readmission. Infection control and hospital epidemiology 2012 Jun;33(6):539–44. PMID: 22561707

20. Fontanarosa PB, McNutt RA. Revisiting Hospital Readmissions. JAMA 2013 Jan 23;309(4):398–400. PMID: 23340644

21. Fox M. Nurse-led early discharge planning for chronic disease reduces hospital readmission rates and all-cause mortality. Evidence-based nursing 2016 Apr;19(2):62. PMID: 26701748

22. Zhou H, Della PR, Roberts P, Goh L, Dhaliwal SS. Utility of models to predict 28-day or 30-day unplanned hospital readmissions: an updated systematic review. BMJ Open 2016 Jun 27;6(6):e011060. PMID: 27354072

23. Naylor MD, Brooten D, Campbell R, Jacobsen BS, Mezey MD, Pauly M V, et al. Comprehensive discharge planning and home follow-up of hospitalized elders: a randomized clinical trial. JAMA 1999 Feb 17;281(7):613–20. PMID: 10029122

24. Koehler BE, Richter KM, Youngblood L, Cohen BA, Prengler ID, Cheng D, et al. Reduction of 30-day postdischarge hospital readmission or emergency department (ED) visit rates in high-risk elderly medical patients through delivery of a targeted care bundle. Journal of hospital medicine 2009 Apr;4(4):211–8. PMID: 19388074

25. Evans RL, Hendricks RD. Evaluating hospital discharge planning: a randomized clinical trial. Medical care 1993 Apr;31(4):358–70. PMID: 8464252

26. Kansagara D, Englander H, Salanitro A, Kagen D, Theobald C, Freeman M, et al. Risk prediction models for hospital readmission: a systematic review. JAMA 2011 Oct 19;306(15):1688–98. PMID: 22009101

27. van Walraven C, Dhalla IA, Bell C, Etchells E, Stiell IG, Zarnke K, et al. Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community. CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne 2010 Apr 6;182(6):551–7. PMID: 20194559

28. Donzé J, Aujesky D, Williams D, Schnipper JL. Potentially avoidable 30-day hospital readmissions in medical patients: derivation and validation of a prediction model. JAMA internal medicine 2013 Apr 22;173(8):632–8. PMID: 23529115

29. Stricker P. Best Practice Strategies and Interventions to Reduce Hospital Readmission Rates [Internet]. 2018. Available from: https://www.tcshealthcare.com/clinical-corner/best-practice-strategies-and-interventions-to-reduce-hospital-readmission-rates/

30. Felix HC, Seaberg B, Bursac Z, Thostenson J, Stewart MK. Why Do Patients Keep Coming Back? Results of a Readmitted Patient Survey. Social Work in Health Care Routledge; 2015;54(1):1–15. PMID: 25588093

31. Hernandez AF, Greiner MA, Fonarow GC, Hammill BG, Heidenreich PA, Yancy CW, et al. Relationship between early physician follow-up and 30-day readmission among Medicare beneficiaries hospitalized for heart failure. JAMA 2010 May 5;303(17):1716–22. PMID: 20442387

32. Report to the Congress: Promoting Greater Efficiency in Medicare. Washington, DC: MedPAC. 2007.

33. Yale New Haven Health Services Corporation/Center for Outcomes Research & Evaluation. 2018 Condition-Specific Measures Updates and Specifications Report Hospital-Level 30-Day Risk-Standardized Readmission Measures. 2018.

34. Zhao P, Yoo I. A Systematic Review of Highly Generalizable Risk Factors for Unplanned 30-Day All-Cause Hospital Readmissions. Journal of Health & Medical Informatics 2017;08(04).

35. Hines AL, Barrett ML, Jiang J, Steiner CA. Conditions With the Largest Number of Adult Hospital Readmissions by Payer, 2011. Healthcare cost and Utilization Project 2014;363(7):1–10. PMID: 24901179

36. Lavenberg JG, Leas B, Umscheid CA, Williams K, Goldmann DR, Kripalani S. Assessing preventability in the quest to reduce hospital readmissions. Journal of Hospital Medicine 2014;9(9):598–603. PMID: 24961204

37. Tanzer MW, Heil EM. Why Majority of Readmission Risk Assessment Tools Fail in Practice. 2013 IEEE International Conference on Healthcare Informatics IEEE; 2013. p. 567–569.

38. Rosen AK, Chen Q, Shin MH, O'Brien W, Shwartz M, Mull HJ, et al. Medical and surgical readmissions in the Veterans Health Administration: what proportion are related to the index hospitalization? Medical care 2014;52(3):243–249. PMID: 24374424

39. Blom MC, Erwander K, Gustafsson L, Landin-Olsson M, Jonsson F, Ivarsson K. The probability of readmission within 30 days of hospital discharge is positively associated with inpatient bed occupancy at discharge--a retrospective cohort study. BMC emergency medicine BMC Emergency Medicine; 2015;15:37. PMID: 26666221

40. Tan SY, Low LL, Yang Y, Lee KH. Applicability of a previously validated readmission predictive index in medical patients in Singapore: a retrospective study. BMC Health Services Research 2013;13(1):366. PMID: 24074454

41. Weeks WB, Lee RE, Wallace AE, West AN, Bagian JP. Do older rural and Urban veterans experience different rates of unplanned readmission to VA and Non-VA hospitals. Journal of Rural Health 2009;25(1):62–69. PMID: 19166563

42. Horkan CM, Purtle SW, Mendu ML, Moromizato T, Gibbons FK, Christopher KB. The Association of Acute Kidney Injury in the Critically Ill and Postdischarge Outcomes: A Cohort Study*. Critical Care Medicine 2015;43(2):354–364. PMID: 25474534

43. van Walraven C, Bennett C, Jennings A, Austin PC, Forster AJ. Proportion of hospital readmissions deemed avoidable: a systematic review. Canadian Medical Association Journal 2011 Apr 19;183(7):E391–E402. PMID: 25517846

44. Benbassat J, Taragin M. Hospital readmissions as a measure of quality of health care: advantages and limitations. Archives of internal medicine 2000 Apr 24;160(8):1074–81. PMID: 10789599

45. Alyahya MS, Hijazi HH, Alshraideh HA, Al-Nasser AD. Using decision trees to explore the association between the length of stay and potentially avoidable readmissions: A retrospective cohort study. Informatics for health & social care 2017 Dec;42(4):361–377. PMID: 28084856

46. National Health Service. Hospital Episode Statistics, Emergency Readmissions to Hospital Within 28 Days of Discharge - Financial year 2011-12 [Internet]. Available from: http://content.digital.nhs.uk/searchcatalogue?productid=13423&q=title%3A%22Emergency+readmissions+to+hospital+within+28+days+of+discharge%22&sort=Relevance&size=10&page=1#top

47. National Health Service. Emergency readmissions within 30 days of discharge

from hospital [Internet]. Available from: http://www.nhs.uk/Scorecard/Pages/IndicatorFacts.aspx?MetricId=8325

48. Dharmarajan K, Hsieh AF, Kulkarni VT, Lin Z, Ross JS, Horwitz LI, et al. Trajectories of risk after hospitalization for heart failure, acute myocardial infarction, or pneumonia: retrospective cohort study. BMJ (Clinical research ed) 2015;350(feb05_19):h411. PMID: 25656852

49. Magdelijns FJH, Schepers L, Pijpers E, Stehouwer CDA, Stassen PM. Unplanned readmissions in younger and older adult patients: the role of healthcare-related adverse events. European journal of medical research BioMed Central; 2016 Sep 15;21(1):35. PMID: 27634174

50. Vest JR, Gamm LD, Oxford BA, Gonzalez MI, Slawson KM. Determinants of preventable readmissions in the United States: a systematic review. Implementation science : IS 2010 Nov 17;5(1):88. PMID: 21083908

51. Berry JG, Ziniel SI, Freeman L, Kaplan W, Antonelli R, Gay J, et al. Hospital readmission and parent perceptions of their child's hospital discharge. International Journal for Quality in Health Care 2013;25(5):573–581. PMID: 23962990

52. Tonkikh O, Shadmi E, Flaks-Manov N, Hoshen M, Balicer RD, Zisberg A. Functional status before and during acute hospitalization and readmission risk identification. Journal of Hospital Medicine 2016;11(9):636–641. PMID: 27130176

53. Mitchell SE, Sadikova E, Jack BW, Paasche-Orlow MK. Health literacy and 30-day postdischarge hospital utilization. Journal of health communication 2012;17 Suppl 3(February 2017):325–338. PMID: 23030580

54. Chakraborty H, Axon RN, Brittingham J, Lyons GR, Cole L, Turley CB. Differences in Hospital Readmission Risk across All Payer Groups in South Carolina. Health Services Research 2017;52(3):1040–1060. PMID: 27678196

55. Lin K-P, Chen P-C, Huang L-Y, Mao H-C, Chan D-CD. Predicting Inpatient Readmission and Outpatient Admission in Elderly: A Population-Based Cohort Study. Medicine 2016;95(16):e3484. PMID: 27100455

56. Lin RJ, Evans AT, Chused AE, Unterbrink ME. Anemia in general medical inpatients prolongs length of stay and increases 30-day unplanned readmission rate. Southern medical journal 2013;106(5):316–320. PMID: 23644640

57. Bisharat N, Handler C, Schwartz N. Readmissions to medical wards: Analysis of demographic and socio-medical factors. European Journal of Internal Medicine European Federation of Internal Medicine.; 2012;23(5):457–460. PMID: 22726376

58. Low LL, Liu N, Wang S, Thumboo J, Ong MEH, Lee KH. Predicting 30-day

readmissions in an Asian population: Building a predictive model by incorporating markers of hospitalization severity. PLoS ONE 2016;11(12):1–16. PMID: 27936053

59. Kim H, Hung WW, Paik MC, Ross JS, Zhao Z, Kim GS, et al. Predictors and outcomes of unplanned readmission to a different hospital. International Journal for Quality in Health Care 2015;27(6):513–519. PMID: 26472739

60. Healthcare Cost and Utilization Project. Clinical Classification Software (CCS) for ICD-9-CM [Internet]. 2015. Available from: https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp

61. Fetter RB, Shin Y, Freeman JL, Averill RF, Thompson JD. Case mix definition by diagnosis-related groups. Medical care 1980 Feb;18(2 Suppl):iii, 1–53. PMID: 7188781

62. Halfon P, Eggli Y, van Melle G, Chevalier J, Wasserfallen JB, Burnand B. Measuring potentially avoidable hospital readmissions. Journal of clinical epidemiology 2002 Jun;55(6):573–87. PMID: 12063099

63. Hasegawa K, Gibo K, Tsugawa Y, Shimada YJ, Camargo CA. Age-Related Differences in the Rate, Timing, and Diagnosis of 30-Day Readmissions in Hospitalized Adults With Asthma Exacerbation. Chest 2016 Apr;149(4):1021–9. PMID: 26836926

64. Dreyer RP, Ranasinghe I, Wang Y, Dharmarajan K, Murugiah K, Nuti S V., et al. Sex Differences in the Rate, Timing, and Principal Diagnoses of 30-Day Readmissions in Younger Patients with Acute Myocardial Infarction. Circulation 2015 Jul 21;132(3):158–66. PMID: 26085455

65. Woz S, Mitchell S, Hesko C, Paasche-Orlow M, Greenwald J, Chetty VK, et al. Gender as risk factor for 30 days post-discharge hospital utilisation: a secondary data analysis. BMJ open 2012 Apr 18;2(2):e000428. PMID: 22514241

66. Rieke K, McGeary C, Schmid KK, Watanabe-Galloway S. Risk Factors for Inpatient Psychiatric Readmission: Are There Gender Differences? Community mental health journal 2016 Aug 25;52(6):675–82. PMID: 26303903

67. Kwan LY, Stratton K, Steinwachs DM, Programs MP, Practice H, Division M. Accounting for Social Risk Factors in Medicare Payment. 2016. PMID: 26844313

68. Baker EH. Socioeconomic Status, Definition. The Wiley Blackwell Encyclopedia of Health, Illness, Behavior, and Society Chichester, UK: John Wiley & Sons, Ltd; 2014. p. 2210–2214.

69. Adler NE, Ostrove JM. Socioeconomic Status and Health: What We Know and What We Don't. Annals New York Academy of Sciences 1999;896(1):3–15. PMID: 10681884

70. Mackenbach JP, Stirbu I, Roskam A-JR, Schaap MM, Menvielle G, Leinsalu M, et

al. Socioeconomic Inequalities in Health in 22 European Countries. INSERM Unité 2008;23(687).

71. Kind AJH, Jencks S, Brock J, Yu M, Bartels C, Ehlenbach W, et al. Neighborhood socioeconomic disadvantage and 30-day rehospitalization: a retrospective cohort study. Annals of internal medicine 2014 Dec 2;161(11):765–74. PMID: 25437404

72. Adler NE, Newman K. Socioeconomic Disparities In Health: Pathways And Policies Inequality. Health Affairs 2002;2(2):60–76.

73. Medicare [Internet]. Available from: https://www.medicare.gov/

74. Joynt KE, Jha AK. A Path Forward on Medicare Readmissions. New England Journal of Medicine 2013 Mar 28;368(13):1175–1177. PMID: 23465068

75. 21st Century Cures Act [Internet]. Available from: https://www.congress.gov/bill/114th-congress/house-bill/34/text

76. Makowsky MD, Klein EY. Identifying the relationship between length of hospital stay and the probability of readmission. Applied Economics Letters 2017 May 9;(410):1–6.

77. Kalra AD, Fisher RS, Axelrod P. Decreased length of stay and cumulative hospitalized days despite increased patient admissions and readmissions in an area of urban poverty. Journal of general internal medicine 2010 Sep 29;25(9):930–5. PMID: 20429040

78. Ritchie C. Health care quality and multimorbidity: the jury is still out. Medical care 2007 Jun;45(6):477–9. PMID: 17515773

79. Fortin M, Soubhi H, Hudon C, Bayliss EA, van den Akker M. Multimorbidity's many challenges. BMJ (Clinical research ed) 2007 May 19;334(7602):1016–7. PMID: 17510108

80. Wolff JL, Starfield B, Anderson G. Prevalence, expenditures, and complications of multiple chronic conditions in the elderly. Archives of internal medicine 2002 Nov 11;162(20):2269–76. PMID: 12418941

81. Valderas JM, Starfield B, Sibbald B, Salisbury C, Rloand M. Defining comorbidity: implications for understanding health and health services. Annals Of Family Medicine 2009;7:357–363. PMID: 19597174

82. Donzé J, Lipsitz S, Bates DW, Schnipper JL. Causes and patterns of readmissions in patients with common comorbidities: retrospective cohort study. BMJ (Clinical research ed) 2013;347(dec16_4):f7171. PMID: 24342737

83. Librero J, Peiró S, Ordiñana R. Chronic comorbidity and outcomes of hospital care: Length of stay, mortality, and readmission at 30 and 365 days. Journal of Clinical Epidemiology 1999;52(3):171–179. PMID: 10210233

84. Zekry D, Loures Valle BH, Graf C, Michel JP, Gold G, Krause KH, et al. Prospective Comparison of 6 Comorbidity Indices as Predictors of 1-Year Post-Hospital Discharge Institutionalization, Readmission, and Mortality in Elderly Individuals. Journal of the American Medical Directors Association Elsevier; 2012;13(3):272–278. PMID: 21450226

85. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. Journal of chronic diseases 1987;40(5):373–83. PMID: 3558716

86. Elixhauser A, Steiner C, Harris DR, Coffey RM. Comorbidity measures for use with administrative data. Medical care 1998 Jan;36(1):8–27. PMID: 9431328

87. Sharabiani MTA, Aylin P, Bottle A. Systematic review of comorbidity indices for administrative data. Medical Care 2012;50(12):1109–1118. PMID: 22929993

88. Deyo RA, Cherkin DC, Ciol MA. Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases. Journal of Clinical Epidemiology 1992;45(6):613–619. PMID: 1607900

89. Leidy NK. Functional status and the forward progress of merry-go-rounds: toward a coherent analytical framework. Nursing research 1994;43(4):196–202. PMID: 8047422

90. Greysen SR, Stijacic Cenzer I, Auerbach AD, Covinsky KE. Functional Impairment and Hospital Readmission in Medicare Seniors. JAMA internal medicine 2015;175(4):559–565. PMID: 24655651

91. Ratzan SC, Parker RM. Health Literacy. In: Ratzan SC, Parker RM, Selden CR, Zorn M, editors. National Library of Medicine Current Bibliographies in Medicine: Health Literacy National Institutes of Health, U.S. Department of Health and Human Services.; 2000. p. v.

92. Health Literacy. Washington, D.C.: National Academies Press; 2004.

93. Evangelista LS, Rasmusson KD, Laramee AS, Barr J, Ammon SE, Dunbar S, et al. Health Literacy and the Patient With Heart Failure-Implications for Patient Care and Research: A Consensus Statement of the Heart Failure Society of America. Journal of Cardiac Failure 2010;16(1):9–16. PMID: 20123313

94. Cloonan P, Wood J, Riley JB. Reducing 30-day readmissions: health literacy strategies. The Journal of nursing administration 2013;43(7–8):382–7. PMID: 23892303

95. Clancy C. An Overview of Measures of Health Literacy. Institute of Medicine (US) Roundtable on Health Literacy Measures of Health Literacy: Workshop Summary Washington, D.C.: National Academies Press (US); 2009.

96. Encinosa WE, Hellinger FJ. The impact of medical errors on ninety-day costs and outcomes: an examination of surgical patients. Health services research 2008

Dec;43(6):2067–85. PMID: 18662169

97. International Classification of Diseases,Ninth Revision, Clinical Modification (ICD-9-CM) [Internet]. Available from: https://www.cdc.gov/nchs/icd/icd9cm.htm

98. Quan H, Sundararajan V, Halfon P, Fong A, Burnand B, Luthi J-C, et al. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. Medical care 2005 Nov;43(11):1130–9. PMID: 16224307

99. Nelder JA, Wedderburn RWM. Generalized Linear Models. Journal of the Royal Statistical Society Series A (General) 1972;135(3):370–384.

100. Cox DR. The Regression Analysis of Binary Sequences. Journal of the Royal Statistical Society Series B (Methodological) 1958;20(2):215–242.

101. Cornfield J. A method of estimating comparative rates from clinical data; applications to cancer of the lung, breast, and cervix. Journal of the National Cancer Institute 1951 Jun;11(6):1269–75. PMID: 14861651

102. Yoo I, Alafaireet P, Marinov M, Pena-Hernandez K, Gopidi R, Chang J-F, et al. Data mining in healthcare and biomedicine: a survey of the literature. Journal of medical systems 2012 Aug;36(4):2431–48. PMID: 21537851

103. Quinlan JR. Discovering Rules by Induction from Large Collections of Examples. Expert Systems in the Micro Electronic Age Edinburgh University Press; 1979.

104. Piatetsky-Shapiro G. Discovery, analysis, and presentation of strong rules. Knowledge Discovery in Databases AAAI/MIT Press; 1991. p. 229–248.

105. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD Statement. European Urology 2015;67(6):1142–1151. PMID: 25561516

106. Krumholz HM, Brindis RG, Brush JE, Cohen DJ, Epstein AJ, Furie K, et al. Standards for statistical models used for public reporting of health outcomes: An American Heart Association scientific statement from the Quality of Care and Outcomes Research Interdisciplinary Writing Group. Circulation 2006;113(3):456–462. PMID: 16365198

107. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 1983 Sep;148(3):839–43. PMID: 6878708

108. Green DM, Swets JA. Signal detection theory and psychophysics. New York, New York, USA: John Wiley & Sons, Inc; 1966.

109. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2 1995. p. 1137–1143.

110. Wang H, Cui Z, Chen Y, Avidan M, Abdallah A Ben, Kronzer A. Predicting Hospital Readmission via Cost-Sensitive Deep Learning. IEEE/ACM transactions on computational biology and bioinformatics 2018;15(6):1968–1978. PMID: 29993930

111. Horne BD, Budge D, Masica AL, Savitz LA, Benuzillo J, Cantu G, et al. Early inpatient calculation of laboratory-based 30-day readmission risk scores empowers clinical risk modification during index hospitalization. American heart journal 2017 Mar;185:101–109. PMID: 28267463

112. Cronin PR, Greenwald JL, Crevensten GC, Chueh HC, Zai AH. Development and Implementation of a Real-Time 30-Day Readmission Predictive Model. AMIA Annu Symp Proc 2014;2014:424–431. PMID: 25954346

113. El Morr C, Ginsburg L, Nam S, Woollard S. Assessing the Performance of a Modified LACE Index (LACE-rt) to Predict Unplanned Readmission After Discharge in a Community Teaching Hospital. Interactive journal of medical research 2017 Mar 8;6(1):e2. PMID: 28274908

114. Cerner Corporation. Health Facts® Database.

115. Yale New Haven Health Services Corporation/Center for Outcomes Research & Evaluation. 2018 All-Cause Hospital Wide Measure Updates and Specifications Report Hospital-Level 30-Day Risk-Standardized Readmission Measures. 2018.

116. International Classification of Diseases,Tenth Revision, Clinical Modification (ICD-10-CM) [Internet]. Available from: https://www.cdc.gov/nchs/icd/icd10cm.htm

117. Healthcare Cost and Utilization Project. Clinical Classification Software (CCS) for ICD10-CM/PCS [Internet]. 2018. Available from: https://www.hcup-us.ahrq.gov/toolssoftware/ccs10/ccs10.jsp

118. International Classification of Diseases,Tenth Revision, Procedure Coding System (ICD-10-PCS) [Internet]. Available from: https://www.cms.gov/Medicare/Coding/ICD10/2020-ICD-10-PCS

119. American Medical Association. Current Procedural Terminology (CPT) [Internet]. Available from: https://www.ama-assn.org/amaone/cpt-current-procedural-terminology

120. The Centers for Medicare & Medicaid Services, America's Health Insurance Plans, Blue Cross and Blue Shield Association. HCPCS Level II codes [Internet]. Available from: https://hcpcs.codes/

121. Szumilas M. Explaining odds ratios. Journal of the Canadian Academy of Child and Adolescent Psychiatry = Journal de l'Academie canadienne de psychiatrie de l'enfant et de l'adolescent 2010 Aug;19(3):227–9. PMID: 20842279

122. Bayes T, Price R. An Essay towards Solving a Problem in the Doctrine of

Chances. By the Late Rev. Mr. Bayes, F. R. S. Communicated by Mr. Price, in a Letter to John Canton, A. M. F. R. S. Philosophical Transactions of the Royal Society of London 1763 Jan 1;53:370–418.

123. Domingos P, Pazzani M. On the Optimality of the Simple Bayesian Classifier under Zero-One Loss. Machine Learning 1997;29:103–130. PMID: 10575050

124. Quinlan JR. Induction of decision trees. Machine Learning 1986;1(1):81–106.

125. Breiman L. Random Forests. Machine Learning 2001;45(1):5–32.

126. Tin Kam Ho. Random decision forests. Proceedings of 3rd International Conference on Document Analysis and Recognition IEEE Comput. Soc. Press; 1995. p. 278–282.

127. Friedman JH. Stochastic gradient boosting. Computational Statistics & Data Analysis 2002 Feb;38(4):367–378.

128. McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. The Bulletin of Mathematical Biophysics 1943 Dec;5(4):115–133.

129. Freund Y, Mason L. The Alternating Decision Tree Algorithm. Proceedings of the 16th International Conference on Machine Learning 1999. p. 124–133.

130. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16 New York, New York, USA: ACM Press; 2016. p. 785–794.

131. Bay SD, Pazzani MJ. Detecting change in categorical data: mining contrast sets. Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '99 New York, New York, USA: ACM Press; 1999. p. 302–306.

132. Emde RN. Risk, intervention and meaning. Psychiatry Elsevier; 1988 Aug;51(3):254–9. PMID: 2464177

133. Turan TN, Al Kasab S, Nizam A, Lynn MJ, Harrell J, Derdeyn CP, et al. Relationship between Risk Factor Control and Compliance with a Lifestyle Modification Program in the Stenting Aggressive Medical Management for Prevention of Recurrent Stroke in Intracranial Stenosis Trial. Journal of stroke and cerebrovascular diseases : the official journal of National Stroke Association 2018 Mar;27(3):801–805. PMID: 29169967

134. Meng X, Brunet A, Turecki G, Liu A, D'Arcy C, Caron J. Risk factor modifications and depression incidence: a 4-year longitudinal Canadian cohort of the Montreal Catchment Area Study. BMJ open 2017 Jun 10;7(6):e015156. PMID: 28601831

135. Chandrasiri A, Dissanayake A, de Silva V. Health promotion in workplaces as a

strategy for modification of risk factors for Non Communicable Diseases (NCDs): A practical example from Sri Lanka. Jayaratne K, De Silva C, Danansuriya M, editors. Work (Reading, Mass) 2016 Oct 17;55(2):281–284. PMID: 27689596

136. Siegel D, Grady D, Browner WS, Hulley SB. Risk factor modification after myocardial infarction. Annals of internal medicine 1988 Aug 1;109(3):213–8. PMID: 3291658

137. Healthcare Cost and Utilization Project. Nationwide Readmissions Database (NRD). Agency for Healthcare Research and Quality, Rockville, MD; 2014.

138. Agrawal R, Srikant R. Fast algorithms for mining association rules. VLDB '94 Proceedings of the 20th International Conference on Very Large Data Bases 1994. p. 487–499.

139. Webb GI. Discovering significant patterns. Machine Learning 2007;68(1):1–33.

140. Fisher RA. On the Interpretation of χ 2 from Contingency Tables, and the Calculation of P. Journal of the Royal Statistical Society 1922 Jan;85(1):87.

141. The United States National Committee on Vital and Health Statistics. Uniform Hospital Discharge Data Set (UHDDS). 1972.

142. Centers for Disease Control and Prevention. National Ambulatory Medical Care Survey Factsheet - Orthopedic Surgery [Internet]. 2010. Available from: https://www.cdc.gov/nchs/data/ahcd/NAMCS_2010_factsheet_orthopedic_surgery.pdf

143. The Centers for Medicare & Medicaid Services. HCAHPS: Patients' Perspectives of Care Survey [Internet]. 2012. Available from: https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/HospitalHCAHPS.html

144. Abtahi AM, Presson AP, Zhang C, Saltzman CL, Tyser AR. Association Between Orthopaedic Outpatient Satisfaction and Non-Modifiable Patient Factors. The Journal of bone and joint surgery American volume 2015 Jul 1;97(13):1041–8. PMID: 26135070

145. Patterson BM, Eskildsen SM, Clement RC, Lin F-C, Olcott CW, Del Gaizo DJ, et al. Patient Satisfaction Is Associated With Time With Provider But Not Clinic Wait Time Among Orthopedic Patients. Orthopedics 2017 Jan 1;40(1):43–48. PMID: 27755644

146. Menendez ME, Loeffler M, Ring D. Patient Satisfaction in an Outpatient Hand Surgery Office: A Comparison of English- and Spanish-Speaking Patients. Quality management in health care 2015;24(4):183–9. PMID: 26426319

147. Tyser AR, Gaffney CJ, Zhang C, Presson AP. The Association of Patient Satisfaction with Pain, Anxiety, and Self-Reported Physical Function. The Journal of bone and joint surgery American volume 2018 Nov 7;100(21):1811–1818.

PMID: 30399075

148. Tisano BK, Nakonezny PA, Gross BS, Martinez JR, Wells JE. Depression and Non-modifiable Patient Factors Associated with Patient Satisfaction in an Academic Orthopaedic Outpatient Clinic: Is it More Than a Provider Issue? Clinical orthopaedics and related research 2019 Aug 20;00:1–9. PMID: 31453885

149. Bible JE, Kay HF, Shau DN, O'Neill KR, Segebarth PB, Devin CJ. What Patient Characteristics Could Potentially Affect Patient Satisfaction Scores During Spine Clinic? Spine 2015 Jul 1;40(13):1039–44. PMID: 25839388

150. Hopkins BS, Patel MR, Yamaguchi JT, Cloney MB, Dahdaleh NS. Predictors of patient satisfaction and survey participation after spine surgery: a retrospective review of 17,853 consecutive spinal patients from a single academic institution. Part 1: Press Ganey. Journal of neurosurgery Spine 2019 Jan 4;30(3):382–388. PMID: 30611140

151. Feagans DR. Patient Satisfaction Analysis in an Outpatient Orthopedic Clinic. University of Missouri; 2018.

152. The Centers for Medicare & Medicaid Services. Technical Assistance Guide for Analyzing Data From the CAHPS® Home and Community-Based Services Survey [Internet]. 2017. Available from: https://www.cms.gov/Research-Statistics-Data-and-Systems/Research/CAHPS/Downloads/HCBS-CAHPS-Data-Analysis.pdf

153. Akaike H. Information Theory and an Extension of the Maximum Likelihood Principle. Selected Papers of Hirotugu Akaike 1998. p. 199–213.

154. Hosmer DW, Lemeshow S. Goodness of fit tests for the multiple logistic regression model. Communications in Statistics - Theory and Methods 1980;9(10):1043–1069.

155. Quinlan JR. C4.5: programs for machine learning. San Francisco, CA: Morgan Kaufmann Publishers Inc.; 1993.

156. Thornton RD, Nurse N, Snavely L, Hackett-Zahler S, Frank K, DiTomasso RA. Influences on patient satisfaction in healthcare centers: a semi-quantitative study over 5 years. BMC health services research BMC Health Services Research; 2017;17(1):361. PMID: 28526039

157. Peck BM. Age-related differences in doctor-patient interaction and patient satisfaction. Current gerontology and geriatrics research 2011;2011:137492. PMID: 22007206

158. Parrish B, Vyas AN, Douglass G. Weighting patient satisfaction factors to inform health care providers of the patient experience in the age of social media consumer sentiment. Patient Experience Journal 2015;2(1):82–92.

159. Voutilainen A, Kvist T, Sherwood PR, Vehviläinen-Julkunen K. A new look at

patient satisfaction: learning from self-organizing maps. Nursing research 2014;63(5):333–45. PMID: 25171559

160. Egloff B, Tausch A, Kohlmann CW, Krohne HW. Relationships between time of day, day of the week, and positive mood: Exploring the role of the mood measure. Motivation and Emotion 1995;19(2):99–110.

161. Husted H, Holm G, Jacobsen S. Predictors of length of stay and patient satisfaction after hip and knee replacement surgery: fast-track experience in 712 patients. Acta orthopaedica 2008 Apr;79(2):168–73. PMID: 18484241

162. Bourne RB, Chesworth BM, Davis AM, Mahomed NN, Charron KDJ. Patient satisfaction after total knee arthroplasty: who is satisfied and who is not? Clinical orthopaedics and related research 2010 Jan;468(1):57–63. PMID: 19844772

163. Kavalnienė R, Deksnyte A, Kasiulevičius V, Šapoka V, Aranauskas R, Aranauskas L. Patient satisfaction with primary healthcare services: are there any links with patients' symptoms of anxiety and depression? BMC family practice 2018;19(1):90. PMID: 29921234

164. Schmocker RK, Cherney Stafford LM, Winslow ER. Disease severity and treatment does not affect satisfaction in diverticulitis. The Journal of surgical research 2017;215:1–5. PMID: 28688633

165. Schmittdiel J, Grumbach K, Selby J V, Quesenberry CP. Effect of physician and patient gender concordance on patient satisfaction and preventive care practices. Journal of general internal medicine 2000 Nov;15(11):761–9. PMID: 11119167

166. Rogo-Gupta LJ, Haunschild C, Altamirano J, Maldonado YA, Fassiotto M. Physician Gender Is Associated with Press Ganey Patient Satisfaction Scores in Outpatient Gynecology. Women's health issues : official publication of the Jacobs Institute of Women's Health 2018;28(3):281–285. PMID: 29429946

167. Derose KP, Hays RD, McCaffrey DF, Baker DW. Does physician gender affect satisfaction of men and women visiting the emergency department? Journal of general internal medicine 2001 Apr;16(4):218–26. PMID: 11318922

VITA

Peng Zhao is a Health Informatics doctoral student at the MU Institute for Data Science and Informatics with a designated graduate minor in Statistics. He has gained a bachelor's degree in Pharmaceutical Sciences from Tianjin University in China and a master's degree in Chemistry from the University of Utah. His research interest spans in the field of health data science and informatics focusing on predictive modeling and exploratory analysis of health outcomes with statistical, data mining, and machine learning methods. He has co-authored several journal articles and one conference article. In his doctoral dissertation project, he applied various predictive modeling and exploratory analysis methods to support the early preventive intervention of unplanned 30-day hospital readmission.