Driving Precision Health Care through Heterogeneous Outcome Analysis

by

Guihua Wang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Business Administration)
in the University of Michigan
2019

Doctoral Committee:

Professor Wallace Hopp, Co-Chair
Assistant Professor Jun Li, Co-Chair
Professor Thomas Buchmueller
Associate Professor Donald Likosky

Guihua Wang

guihuaw@umich.edu

ORCID iD: 0000-0001-8324-7073

Acknowledgments

I am so blessed to have Professors Wallace Hopp and Jun Li as my advisors and dissertation committee co-chairs. Their commitment and passion for high-quality and impactful research has been one of the biggest inspirations in my life. I want to thank them for transforming me from a student to a professor and helping me realize my dream of becoming a faculty member. I also want to thank my other committee members, Professors Thomas Buchmueller and Donald Likosky, for their time, support, and guidance throughout the preparation and review of this dissertation.

The faculty members of the Technology and Operations group at Michigan Ross provided me lots of feedback and suggestions on my research, teaching, and career. In particular, our department chairs, Izak Duenyas and Roman Kapuscinski, and PhD coordinators, Stephen Leider, Amitabh Singha, and Joline Uichanco, met me regularly to check on my academic progress and guided me towards meeting the dissertation requirements. My course co-instructors Stefanus Jasin and Mohamed Mostagir helped me improve my teaching and presentation skills. Other faculty members including Hyun-Soo Ahn, Ravi Anupindi, Damian Beil, Samantha Keppler, Bill Lovejoy, Shima Nassiri, John Silberholz, and Andrew Wu helped me with my academic presentations and interviews.

My PhD journey would not have been so enjoyable without the accompanying of my classmates and friends. My cohort mates Behrooz, Dan, Lai, Longxiu, and Yifei organized many group lunches and parties to help us maintain student-life balance. My seniors Anyan, Evgeny, George, Iris, Murray, and Yao generously shared their past experience of being a PhD student and offered valuable advice on finding an academic job. My juniors Aravind, Feng, Jiaxin, Miao, Samer, Yu, and Zoey showed their genuine interest in my research and provided candid feedback on my mock interviews and job talks related to this dissertation.

Finally, I want to thank my parents and inlaws for their unconditional love and support. A very special thanks to my sister and brothers-in-law for taking care of our parents and grandparents while my wife and I were not around. I am indebted to my wonderful wife for encouraging me to find my interest and pursue a career in academia. She has taken more than her fair share of the work to keep our family functioning. Thanks to our son Jerry for making me a daddy – you made me feel that I might have accomplished more than finishing a dissertation.

TABLE OF CONTENTS

Acknowledgments . . . . **ii**

List of Figures . . . . **v**

List of Tables . . . . **vi**

List of Appendices . . . . **viii**

Abstract . . . . **ix**

Chapter

1 Introduction **1**

2 Using Patient-Specific Quality Information to Unlock Hidden Health Care Capabilities **3**

2.1 Motivation 3
2.2 Literature Review 5
2.3 Empirical Setting and Data 7
2.3.1 Mitral Valve Disease 8
2.3.2 Data Description 9
2.3.3 Data Preparation 9
2.3.4 Measures of Provider Quality 10
2.4 Empirical Model of Provider Quality 11
2.4.1 Factors Affecting Surgical Outcomes 11
2.4.2 Quality Model 12
2.4.3 Estimation of Hospital and Surgeon Effects 14
2.5 Generating Patient-Specific Quality Information 15
2.5.1 Summary Statistics and Estimation Results 16
2.5.2 Population-Average vs. Patient-Specific Outcomes 17
2.5.3 Quality-Adjusted Life Expectancy 20
2.6 Estimating the Value of Patient-Specific Information 23
2.6.1 Patient Choice Model 24
2.6.2 Simulation Results 26
2.6.3 Generalizing Our Results 30
2.7 Conclusion and Managerial Implications 31

3 An Instrumental Variable Tree Approach for Detecting Heterogeneous Treatment Effects in Observational Studies **34**

3.1 Introduction 34
3.2 Literature Review 37
3.3 Problem Formulation and the IV Tree Approach 38
3.3.1 Regression Approach 39
3.3.2 Tree-Based Approaches 40
3.3.3 The IV Tree Approach 41
3.3.4 IV Tree Algorithm 43
3.3.5 Asymptotic Properties 46
3.4 Performance on Synthetic Data 49
3.4.1 IV Tree Performance and Comparison with Causal Tree 49
3.4.2 Comparison with the Generalized Random Forest 54
3.5 Empirical Example: Teaching vs. Non-Teaching Hospitals 57
3.5.1 Data Description, Outcome Measure, and Feature Space . . . 58
3.5.2 Instrumental Variable Construction 59
3.5.3 Results and Discussion 61
3.6 Conclusion 63

4 Personalized Health Care Outcome Analysis of Cardiovascular Surgical Procedures 66

4.1 Introduction 66
4.2 Literature Review 70
4.3 The Empirical Model 73
4.3.1 Problem Description 73
4.3.2 Instrumental Variable Tree Approach 74
4.4 Empirical Setting and Data 79
4.4.1 Data Description and Preparation 80
4.4.2 Outcome Measure and Feature Space 80
4.4.3 Instrumental Variable Method 83
4.5 Results and Discussion 83
4.6 Managerial Implications 86
4.6.1 Implications for Patients 87
4.6.2 Implications for Hospitals and Payers 89
4.7 Conclusion 95

Appendices 97

Bibliography 112

LIST OF FIGURES

Figure

2.1 Mitral Valve Repair Rate by Surgeon for A Patient with Average Characteristics . . . 18
2.2 Mitral Valve Repair Rate by Surgeon for Patients of Different Levels of Acuity . . . 19
2.3 Related Events and Decision Process for Patients with Mitral Valve Diseases 21
2.4 Difference in Quality-Adjusted Life Expectancy (Months) between 95th-percentile and 50th-percentile Surgeons . . . 22

3.1 Outcome Differences between Teaching and Non-Teaching Hospitals by Patient Subgroup . . . 65

4.1 Comparison of Complication Score for Different Patient Groups (IV Tree) 86
4.2 Complication Reduction by Patient Group . . . 91
4.3 Complication Reduction under Patient-Centric Information . . . 91
4.4 Comparison of Hospitals' Performance by Patient Group . . . 93
4.5 Percentage Payment under Patient-Centric and Population-Average Measure 94

A.1 Complication Rate by Surgeon for A Patient with Average Characteristics 99
A.2 Readmission Rate by Surgeon for A Patient with Average Characteristics 99
A.3 Complication Rate by Surgeon for Patients of Different Levels of Acuity . 100
A.4 Readmission Rate by Surgeon for Patients of Different Levels of Acuity . . 100

C.1 Rank of the Chosen Hospital by Travel Distance . . . 111

LIST OF TABLES

Table

2.1 Comparison of Mitral Valve Repair and Replacement 8
2.2 Summary of Patient Characteristics 16
2.3 Estimation Results of the Quality Model 17
2.4 Comparison of the Effectiveness of Patient-Specific Information and Capacity Increase 27
2.5 Comparison of the Values from Using Patient-Specific and Population-Average Information 29

3.1 Designs of the Simulation Study 50
3.2 Comparison of IV Tree (IVT) and Causal Tree (CT) 52
3.3 Mean-Squared Errors of IV Forest (IVF) and Generalized Random Forest (GRF) 55
3.4 Split Frequencies of IV Forest (IVF) and Generalized Random Forest (GRF) in Design 2 56
3.5 Severity Score and Incidence Rate of Different Complications 60
3.6 Relationship between Distance to Teaching Hospital and Patient Characteristics 62
3.7 Relationship between Expected Teaching and Patient Characteristics . . . 62
3.8 Comparison of IV Tree (IVT) and Causal Tree (CT) by Statistical Significance 63
3.9 Comparison of IV Tree (IVT) and Causal Tree (CT) for Different Patients 64

4.1 Relative Performance of Hospitals for Different Procedures 67
4.2 Weights of Different Complications 82
4.3 Summary of Complication Score for Different Procedures 82
4.4 Comparison of Complication Score with the State Average for Different Patients 85
4.5 Complication Reduction from Using Patient-Centric Instead of Population-Average Information in Hospital Selection 90

A.1 Hausman Test Results 98
A.2 Estimation Results of Each Type of Complication 101
A.3 Comparison of the Effectiveness of Patient-Specific Information and Capacity Increase 104
A.4 Comparison of the Values from Using Patient-Specific and Population-Average Information 104

C.1 Relationship between Distance to the Nearest Hospital and Patient Characteristics 110

LIST OF APPENDICES

A Appendix to Chapter 2 97

B Appendix to Chapter 3 105

C Appendix to Chapter 4 110

ABSTRACT

This dissertation is based on three essays that examine how to measure the heterogeneity of patient outcomes using readily available data, how to use the results to generate patient-centric outcome information, and how outcome data can be used to benefit patients, payers, and providers. In the first essay, we document a wide variation in quality among 188 surgeons at 35 hospitals in New York State that perform mitral valve surgery. Our analysis shows that patients of different demographics and levels of acuity benefit differently from elite surgeons. We estimate that the total societal benefits from using our proposed patient-centric information are comparable to those achievable by enabling the best surgeons to treat 10%–20% more patients under currently available population-average information. In the second essay, we develop a technique that incorporates the instrumental variable method into a causal tree to correct for potential endogeneity biases in heterogeneous treatment effect analysis using observational data. The resulting instrumental variable tree (IV tree) approach partitions subjects into subgroups with similar treatment effects within subgroups and different treatment effects across subgroups. In the third essay, we provide empirical evidence that outcome differences between health care providers are heterogeneous across different patients. We then use the IV tree approach to identify patient types that exhibit significant differences in outcome quality. After that, we quantify the differences in patient outcomes between providers in a (patient-centric) manner that is useful to individual patients. Lastly, we show that providing patient-centric outcome information not only helps patients choose providers but also helps providers identify areas for improvement and payers design cost-effective payment.

CHAPTER 1

Introduction

Patients differ not only in their demographics and medical conditions but also in their responses to a treatment. While some patients respond positively to a particular treatment, others may see little response or even experience serious negative effects from the same treatment. However, most studies of treatment effects have focused on the average effect across all patients in a sample. The lack of a large sample and an effective methodology for analyzing heterogeneous treatment effects has led to "one-size-fits-all" approaches that may not help, or may even harm, some patients.

The growing availability of observational data provides a unique opportunity for personalize health care outcome analysis. However, there are two main challenges of using big observational data for heterogeneous treatment effect analysis. First, the high-dimensionality of big data makes it unclear how to partition patients into subgroups with similar treatment effects within subgroups and different treatment effects across subgroups. Second, the uncontrolled nature of observational data introduces potential endogeneity issues because the data often do not include all features that affect treatment assignment and outcome. We consider health care providers as different treatments and address both challenges.

In the first study, we document a wide variation in quality among 188 surgeons at 35 hospitals in New York State who perform mitral valve surgery. Our analysis shows that patients of different demographics and levels of acuity benefit differently from elite surgeons. However, existing healthcare provider quality information is based on population averages, so it does not differentiate patients of different medical conditions. This implies that patient-centric quality information, which calibrates outcome statistics by patient demographics and acuity, can increase the ability of patients to choose the most appropriate surgeon. In this study, we develop an econometric model for computing patient-centric information from outcome data and evaluate the potential health benefits from using such information to guide patients to surgeons. We estimate that the total societal benefits from using patient-centric information are comparable to those achievable by enabling the best surgeons to treat 10%–20% more patients under population-average information.

In the second study, we develop a technique that incorporates the instrumental variable method into a causal tree to correct for potential endogeneity biases in heterogeneous treatment effect analysis using observational data. The resulting instrumental variable tree approach partitions subjects into subgroups with similar treatment effects within subgroups and different treatment effects across subgroups. The estimated treatment effects are asymptotically consistent under very general assumptions. Using simulated data, we show that our approach has better coverage rates and smaller mean-squared errors than the conventional causal tree, and that a forest constructed using instrumental variable trees has better accuracy and interpretability than the generalized random forest.

In the third study, we focus on six cardiovascular surgical procedures and provide empirical evidence that outcome differences between health care providers are heterogeneous across different patients. We then use the instrumental variable tree approach to identify patient types that exhibit significant differences in outcome quality. After that, we quantify the differences in patient outcomes between providers in a (patient-centric) manner that is useful to individual patients. Lastly, we show that providing patient-centric outcome information not only helps patients choose providers but also helps providers identify areas for improvement and payers design cost-effective payment programs.

CHAPTER 2

Using Patient-Specific Quality Information to Unlock Hidden Health Care Capabilities

2.1 Motivation

How to “fix health care” is one of the most hotly debated topics in all of American society. Academic articles, media programs, legislative debates and water cooler conversations are rife with recommendations on how to provide patients with better and more cost effective health care. The vast majority of these proposals, from reimbursement bundling and accountable care organizations to patient care paths and lean transformations, are aimed at changing health care delivery and/or payment structure. However, a widely overlooked reality is that better health care is available right now within the American system. But too often patients can’t find it.

To understand what is possible and how it might be achieved, we delve into the details of a specific medical condition at the surgeon level by focusing specifically on patients with mitral valve disease and addressing two key questions: (1) how do different types of patients benefit differently from elite surgeons (i.e., those perform significantly better than the state average), and (2) How can outcome data be used to improve health care at both the individual and societal levels? Unfortunately, while simple to state, these questions are not straightforward to analyze using currently available data.

To answer the first question, we need to characterize the performance of the surgeons that treat mitral valve patients. There are various consumer-oriented healthcare rating systems that attempt to do this by providing provider quality information. For example, the Centers for Medicare and Medicaid Services maintain the Hospital Compare website that reports on over 4,000 Medicare-certified hospitals across the country with regard to quality of care, safety measures and patient satisfaction. Various non-profit organizations, including the Leapfrog Group, Consumer Reports and the California

Health Care Foundation, and private companies such as US News and Healthgrades, also share self-reported hospital quality information and rankings via websites.

While these sources provide useful information, they fall well short of providing the data mitral valve patients need to make accurate comparisons of providers because: (1) they generally aggregate ratings into broad categories such as heart surgery, rather than reporting them for individual procedures such as mitral valve surgery; (2) most of these analyses are performed at the hospital level and do not provide information about individual surgeons within a hospital; and (3) most ratings do not indicate the magnitude of the difference between levels (e.g., between "1-star" and "3-star", or between "average" and "above average").

Some states, such as New York, address these issues by compiling risk-adjusted mortality rates for individual hospitals and surgeons that perform CABG and/or mitral valve surgery.[1] They also indicate whether a hospital or a surgeon is statistically significantly better or worse than the state average. However, there are still some remaining issues. First, they focus primarily on mortality rates (and sometimes on complication and readmission rates), where the low probability of events can make it difficult to discern statistical differences among providers. When all or most providers are not statistically different from the state average, patients don't have a basis for identifying better health care. However, if we focus on a specific medical condition, there may be measures of quality that show more variation across providers. In this study, we focus on mitral valve surgery and introduce mitral valve repair rate to the conventional quality metrics (i.e., rates of mortality, complication and readmission). As we will show later in the paper, repair rate is an informative measure of surgeon quality and patient benefit, and a metric that shows significant variation among providers.

Second, all existing quality ratings are based on population-average outcome measures,[2] and therefore do not provide personalized guidance to patients of different demographics and levels of acuity. Population-average quality information has two major issues. First, patients care more about quality information specific to their procedure types and medical conditions than about information about an average patient who may not even exist. Second, population-average information suggests that all patients will benefit equally from an elite surgeon. However, it isn't reasonable to expect all

[1] The NY state cardiac surgery reporting system includes patient demographics (age, gender), comorbidities (e.g., lung disease, diabetes and renal failure), previous procedures, hemodynamic state, ventricular function and vessel diseased as control variables in the model. See https://www.health.ny.gov/statistics/diseases/cardiovascular for more details.

[2] Population-average outcome measures refer to those focusing on the "average effect" or "homogeneous effect" of a given treatment (Kravitz et al. 2004).

patients to go to the single best surgeon, since this would create an impossible capacity imbalance. But, with patient-specific quality information, as we will show for the case of mitral valve surgery, it is possible to achieve substantial improvements in outcomes without overloading any single surgeon.

To address the second question of how outcome data can be used to improve health care, we examine two potential approaches for leveraging patient outcome data: (1) using population-average information, possibly accompanied by measures to increase surgeon capacity so that more patients can be treated by the best surgeons, and (2) providing patient-specific information so that patients can better balance benefits with costs of traveling/waiting to see an elite surgeon for treatment. The second approach is motivated by the fact that existing quality information is almost always based on population averages and so does not indicate differences in patient benefits from elite surgeons. Consequently, patients who are treated by these surgeons may not be those who benefit the most. If patients, and the cardiologists who refer them, have access to patient-specific quality information and use it in selecting a surgeon, the patients with the most to gain will be those most inclined to incur additional costs associated with being treated by an elite surgeon. We will make use of empirical and analytical models to estimate how much the resulting alignment of patients with surgeons improves aggregate patient welfare.

Our work makes two contributions to existing studies of provider quality and health care quality information. First, we find that the quality gap (e.g., difference in repair rate) between surgeons is heterogeneous for different patients. For the same level of increase in repair rate, younger patients have more years to live and healthier patients have a higher quality of life. To account for such differences, we create a model to calculate quality-adjusted life expectancy that combines many of short- and long-term effects of the single number quality metrics. Second, we construct a patient choice model to compare scenarios where patients choose providers based on population-average or patient-specific information. We find that providing patient-specific information helps patients to find better care. The societal benefits from using patient-specific information are comparable to those achievable by enabling the best surgeons to treat 10–20% more patients under population-average information.

2.2 Literature Review

There has been growing interest in studying hospital quality since 1989 when the Agency for Health Care Policy and Research was created by Congress in response to a report of wide geographic variations in practice patterns among hospitals in the US (see e.g., Chassin et al. 1987). In a seminal paper, Keeler et al. (1992) compared 297 US hospitals for congestive heart failure, acute myocardial infarction, pneumonia, stroke or hip replacement, and found that quality varied from state to state, but that quality was generally better in teaching, large, and urban hospitals than in non-teaching, small, and rural hospitals. Subsequent studies have also found that high-volume hospitals tend to perform better than low-volume hospitals (Birkmeyer et al. 2002, Gammie et al. 2009, Vassileva et al. 2012) and that high-volume surgeons tend to perform better than low-volume surgeons (Birkmeyer et al. 2003, Bolling et al. 2010, Kilic et al. 2013). Instead of studying hospital characteristics that are associated with performance, a number of studies looked directly at hospital fixed effects and found similar quality variations across hospitals (Lingsma et al. 2010, McClellan et al. 1994, Moran et al. 2014). Note that, in these studies, the effects of hospital volume and other characteristics are absorbed by hospital fixed effects. By using a multilevel model, we are able to separate out volume effects and unobservable hospital specific effects beyond those captured by volume.

In the operations management literature, a number of studies have examined factors affecting health care quality. Some of these have focused on surgeon experience and its impact on surgical outcome. For example, KC and Staats (2012) investigated the differential effects of focal and related experience, and found that surgeon focal experience has a greater effect than related experience on surgeon performance. KC et al. (2013) examined how surgeons learn from their own and others' experiences, and found that individuals learn more from their own successes but also from others' failures. Ramdas et al. (2017) studied how learning and forgetting affect surgical outcomes by analyzing a surgeon's experience with specific surgical device versions and the time between their repeated uses. Other studies have analyzed the impact of workload on quality and patient outcome. For instance, Kim et al. (2014) examined the impact of ICU congestion on a patient's care pathway and the subsequent effect on patient outcomes, and found that the impact of ICU admission is highly variable for different patients and different outcomes. Jaeker and Tucker (2016) studied the relationship between workload and patient length of stay (LOS), and found that the effects of inpatient workload on LOS propagate across patient types. Freeman et al. (2016) show that gatekeeper providers (midwives in their study) ration resource-intensive discretionary services and also increase the rate of specialist referrals when workload increases. In addition to surgeon experience and workload, queue management (Song et al. 2015) and secure messaging between patients and physicians via patient portals (Bavafa et al. 2018) have also been found to affect productivity and patient outcome. However, none of these studies have compared quality among health care providers or studied the impact of patient-specific information on outcomes.

The findings from this line of research suggest that: (1) both hospitals and surgeons play pivotal roles in determining healthcare quality; (2) experience at both institutional and individual levels significantly affects quality; and (3) in addition to experience, many other nuanced factors affect provider quality.

We contribute to the literature of healthcare provider quality evaluation by incorporating hospital and surgeon volume effects, as well as hospital and surgeon specific effects, in our model. However, unlike the aforementioned studies, our focus is not to identify the effect of a specific factor on provider quality, but rather to examine the quality gap among providers, which allows us to offer insights on how to best utilize the capabilities of existing healthcare system.

To accomplish this, we first need to identify elite surgeons who produce better quality outcomes. More importantly, we need to quantify the quality gap between an elite surgeon and an average surgeon for patients of different demographics and levels of acuity. Most prior studies have focused on measuring provider quality based on the population average (i.e., risk adjusted) outcomes, and thereby assume away heterogeneity in outcomes among patients of different demographics and levels of acuity. As a result, their assessments of provider quality apply to average patients (who may not even exist) but may not be useful for a given patient. Recognizing this flaw in population average information, a number of observers have called for a patient-centered focus in both patient care and in quality assessment (see, e.g., FDA 2013, Gerteis 1993, IOM 2011, Kattan and Vickers 2004, Kent and Hayward 2007, Kravitz et al. 2004). Our study contributes to the literature on patient-centered care by proposing patient-specific quality information as a means to help patients find care best suited to their needs.

2.3 Empirical Setting and Data

We choose mitral valve surgery as the empirical setting for our analysis of health care provider quality for several reasons. First, mitral valve disease is the most common form of heart valve disease in US. It affects 5% of the population and results in over 500,000 hospital admissions per year.[3] Second, mitral valve repair is a relatively new and complicated procedure. Because of the high level of skill required, surgeons may

[3] http://heartvalvedisease.nm.org/mitral-valve-disease.html

differ substantially in their outcomes. Third, there are many extant medical studies that provide data on the clinical outcomes of treatments available to mitral valve patients.

2.3.1 Mitral Valve Disease

The mitral valve is located between the left chambers of the heart. Its main function is to allow blood to flow from the left atrium to the left ventricle but not in the other direction. Mitral valve disease refers to conditions that compromise the ability of the mitral valve to seal against the backflow of blood.

There are two clinical options for the correction of mitral valve disease — mitral valve repair and mitral valve replacement. Mitral valve repair restores the function of the original valve, and is therefore the preferred option (Bolling et al. 2010). Table 2.1 compares the risks of mortality and complications associated with both procedures for a 60-year old male patient without major comorbidities (Society of Thoracic Surgeons, 2016). We see that the risks associated with replacement are 44.8% to 94.3% higher than those associated with repair.

Table 2.1: Comparison of Mitral Valve Repair and Replacement

	Mitral Valve Repair	Mitral Valve Replacement	Relative Gap
Operative Mortality	0.4%	0.7%	94.3%
Prolonged Ventilation	2.7%	5.1%	85.8%
Renal Failure	0.9%	1.5%	63.5%
Reoperation	4.7%	6.8%	44.8%

Source: Society of Thoracic Surgeons Risk Evaluator (2016).

Consequently, surgeons strive to repair a mitral valve whenever possible. However, since it is impossible to guarantee a repair, surgeons always have either a biological valve (from a cow or pig) or a mechanical valve (made of special carbon compounds and titanium) ready as a backup. If, during the procedure, visual inspection reveals the valve is not repairable, or a repair is attempted but fails (e.g., leaks), a replacement valve will be installed. The likelihood of a repair depends on both patient characteristics and surgeon skill. Hence, repair rate (fraction of patients whose valves are repaired) is an indicator of surgeon quality after controlling for the mix of patients.

2.3.2 Data Description

We used data from New York State that describe 10 million in- and out-patient discharges from all hospitals in New York from 2009–2012. These data contain patient-level clinical and resource-use information, including admission status (e.g., elective, emergent and urgent), patient demographics and comorbidities, hospital and physician identifiers, and principal and secondary diagnoses. For each discharge, the data record whether a patient received a mitral valve repair or replacement. They also indicate whether a patient died or experienced other complications during the procedure or post-surgery hospitalization. Because they record all visits, we are able to identify readmissions for the same patient across hospitals and time. Finally, the data include 5-digit zip codes of patients' home and hospital addresses, which allow us to estimate travel distance from each patient's home to any hospital.

2.3.3 Data Preparation

We identified discharges related to mitral valve surgery by using the clinical codes 35.12, 35.23 and 35.24 in the International Classification of Disease (9th revision). To focus on isolated mitral valve surgery, we followed previous studies (e.g., Vassileva et al. 2012) in excluding patients who were less than 30 years old, had coronary revascularization, congenital heart disease, excision of ventricular aneurysm, replacement of thoracic aorta, aortic fenestration procedure, closed heart valvuloplasty, heart transplant, or other valvular repair.

Because the ultimate objective of this study is to allow patients to choose the most appropriate care for them, we focused on elective patients only, as opposed to emergent or urgent patients whose choice of providers may have been constrained by the urgency of their medical condition (Batt and Terwiesch 2015). An elective mitral valve patient can wait for a year or more from diagnosis to treatment (Carroll et al. 1995), which allows for considerable flexibility in the choice of providers.

Lastly, we focused on New York patients who were treated in New York hospitals. We do not directly observe New York residents who were treated outside New York because we lack data to compute patient-to-hospital distances for these patients.[4] This is unlikely to cause a sampling concern in our context, because New York has 4 out of the 50 nationally ranked heart programs.[5] If a patient decides to seek a better

[4] Although many others states in the US make their inpatient and outpatient discharge data available, most do not contain patient-level zip code information without which we cannot estimate distances to out-of-state providers.

[5] http://health.usnews.com/best-hospitals/rankings/cardiology-and-heart-surgery

provider than those available locally, the best providers in New York are comparable to the best providers in the country. We also excluded patients who traveled from other states to New York for mitral valve surgery, because we do not have sufficient data on out-of-state hospitals to describe these patients' local treatment options.[6] We believe this exclusion also does not bias the estimation of provider quality, because it is very unlikely that provider quality varies for in-state vs. out-of-state patients conditional on patient demographics and medical conditions.

2.3.4 Measures of Provider Quality

We use the rates of mortality, complication, readmission and mitral valve repair as measures of quality for two main reasons. First, mortality, complication and readmission rates are commonly used measures of health care provider quality in both hospital rating systems (e.g., US News, the Leapfrog Group, Healthgrades, Hospital Compare, and New York State Cardiac Surgery Reporting System) and existing literature (e.g., Birkmeyer et al. 2002, Brooks et al. 2006, Gupta et al. 2014). They describe important short-term risks to patients while they are staying in hospitals and the time shortly after discharge. Because these quality metrics capture different aspects of short-term risks patients care about, we include all three of them in this study. To capture the long-term risks, we follow existing medical literature by including mitral valve repair rate as an additional quality metric (see e.g., Vassileva et al. 2012). The inclusion of multiple metrics ensures that the conclusions we draw from this study is not driven by the choice of a specific quality metric.

Second, the above metrics provide the basis for constructing a single metric that captures benefits from visiting elite surgeons. In Section 2.5.3, we describe a quality-adjusted life expectancy metric that combines the main short- and long-term effects of the single number metrics. This allows us to incorporate the heterogeneity in multiple aspects of provider quality, capture different aspects of patient benefits, and assess the value of patient-specific information.

In this study, the quality measures are operationalized as follows. Mortality is measured as death during hospitalization.[7] Complication is measured as occurrence of one or multiple mitral valve related complications including stroke, wound infection, renal failure, reoperation and ventilation observed during hospitalization (Society of

[6] Among all the patients treated at NY hospitals, around 90% of them were from NY. Most of the other 10% of patients came from nearby states such as New Jersey (7%).

[7] The New York in- and out-patient discharge data do not track post-discharge death or complication.

Thoracic Surgeons 2016). To analyze readmission rate, we focus on 30-day readmission by identifying patients who were admitted to the same or other hospitals within 30 days after discharge.[8] Lastly, we observe from the data whether a patient received mitral valve repair or replacement based on the clinical procedure codes.

2.4 Empirical Model of Provider Quality

While it is widely accepted that health care providers differ with regard to quality, it is not clear whether the outcome differences between providers are heterogeneous across patients of different demographics and levels of acuity, and if so, to what extent. In this section, we construct an econometric model of provider quality to evaluate how different types of patients benefit differently from elite surgeons.

2.4.1 Factors Affecting Surgical Outcomes

Surgical outcomes can be affected by patient, hospital and surgeon characteristics. For example, old age correlates with increased risks of mortality, complication and readmission (Gupta et al. 2014, Merkow et al. 2015, Society of Thoracic Surgeons 2016). Studies have found that white patients are less likely to have complications and unplanned readmissions than are black and Hispanic patients, and that female patients are more likely than male patients to have these undesired events (Iribarne et al. 2014, Merkow et al. 2015, Society of Thoracic Surgeons 2016). Such differences can be the result of medical (e.g., comorbidities) or behavioral (e.g., delay in undergoing cardiac surgery) differences between various patient groups (Fasken et al. 2001). Comorbidities that increase the risk of mortality, complication and readmission include diabetes, chronic obstructive pulmonary disease, hypertension and renal failure (Gupta et al. 2014, Iribarne et al. 2014, Merkow et al. 2015, Society of Thoracic Surgeons 2016).

With respect to repair rate in mitral valve surgery, Bolling et al. (2010) and Vassileva et al. (2013) separately found that younger and white patients are more likely to receive a repair, whereas females are less likely to receive a repair. Presence of various comorbidities including atrial fibrillation, chronic obstructive pulmonary disease, diabetes, heart failure, renal disease and hypertension also reduces the likelihood of mitral valve repair (Daneshmand et al. 2009, Savage et al. 2003, and Vassileva et al. 2013).

[8] https://www.cms.gov/medicare/medicare-fee-for-service-payment/acuteinpatientpps/readmissions-reduction-program.html

Of more direct interest to us in this study is the impact of hospitals and surgeons on surgical outcomes. Presumably a hospital with more skilled surgeons, as well as more experienced support teams and an organizational structure that promotes learning and quality improvement, will have better quality than a hospital without these assets. However, because quality and its antecedents are challenging to measure, patients and researchers alike must often rely on proxies to gauge the provider (hospital and/or surgeon) effect on quality. One of the most common proxies is surgical volume (Birkmeyer et al. 2002, Birkmeyer et al. 2003, Gammie et al. 2009, Kilic et al. 2013, Vassileva et al. 2012, Vassileve et al. 2013).

Note, however, volume alone is not sufficient to capture all variations in surgeon skills or hospital effects. Some surgeons may have better training or higher innate ability. At the hospital level, initiatives focused on quality improvement have also proven to be very effective (Barr et al. 2006, Lindenauer et al. 2007). Therefore, it is imperative to account for variations in surgeon and hospital quality beyond those captured by surgical volume.

2.4.2 Quality Model

To evaluate the impact of surgeon and hospital on patient outcomes, we need an econometric model that can address the following three issues. First, our data has a nested structure, i.e., patients are grouped under different surgeons and surgeons are grouped under different hospitals. Second, outcomes of patients treated by the same surgeon/hospital may correlate with each other due to unobservable surgeon/hospital characteristics. Third, we want to separately identify surgeon and hospital effects as well as their volume effects. To address these issues, we follow the approaches of Healthcare Cost and Utilization Project[9] and Centers for Medicare and Medicaid Services[10] and use a multilevel probit model (Gibbons and Hedeker 1997).

Let Y^*_{ijk} denote the latent variable associated with the outcome measure (i.e., mortality, complication, readmission, or repair) of patient i treated by surgeon j at hospital k. Y^*_{ijk} can be measured as a function of patient, surgeon and hospital characteristics:

$$\begin{aligned} Y^*_{ijk} &= \gamma_0 + \gamma_1 Age_i + \gamma_2 Gender_i + \gamma_3 Race_i + \gamma_4 Comorb_i \\ &\quad + \gamma_5 SurgVol_i + \gamma_6 HospVol_i + \alpha_k + \beta_{jk} + \epsilon_{ijk} \\ Y_{ijk} &= \mathbf{1}\{Y^*_{ijk} > 0\} \end{aligned} \tag{2.1}$$

[9] https://www.hcup-us.ahrq.gov/reports/methods/2007_01.pdf

[10] https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/Measure-Methodology.html

where $SurgVol_i$ and $HospVol_i$ are measures of surgeon and hospital volumes of the procedure received by patient i,[11] α_k represents the unobserved effect of hospital k, and β_{jk} represents the unobserved effect of surgeon j at hospital k.[12] We assume these unobserved effects are drawn from two normal distributions:

$$\alpha_k \sim N(\mu_\alpha, \sigma_\alpha^2), \beta_{jk} \sim N(\mu_\beta, \sigma_\beta^2)$$

where μ_α and μ_β represent the mean hospital and surgeon effects, and σ_α^2 and σ_β^2 represent between-hospital and between-surgeon variations after accounting for hospital volume, surgeon volume and patient conditions at admission. If there are no between-hospital or between-surgeon differences in the outcomes beyond those captured by surgical volume and patient characteristics, then $\sigma_\alpha^2 = 0$ (i.e., $\alpha_1 = \alpha_2 = ... = \alpha_K$) or $\sigma_\beta^2 = 0$ (i.e., $\beta_1 = \beta_2 = ... = \beta_J$).

For all models, we have robust standard errors clustered by surgeon to allow for differences in the variance/standard errors due to arbitrary intra-group correlation (KC and Terwiesch 2011, Jaeker and Tucker 2016).

The model specified above is essentially a random effects model. Alternatively, one could also specify a fixed effects model at the surgeon level. There are two approaches to account for surgeon fixed effects. One approach is to include dummies for surgeons. This approach estimates surgeon fixed effects explicitly. In our setting, however, this would require a total of 187 surgeon dummies, which would create a collinearity issue preventing us from identifying the fixed effects of 75 surgeons. Moreover, many surgeon fixed effects would not be able to be estimated reliably because over 50% of surgeons performed fewer than 10 cases during our study period. The other approach to account for surgeon fixed effects is to specify them as unobserved error terms. In particular, this could be done through fixed effects logit model (Wooldridge 2010, Chapter 15.8), which does not treat fixed effects as parameters to estimate and uses only within-subject variations for estimation. Also, it allows for any association between the unobserved surgeon fixed effects and patient characteristics (observed and unobserved). However, because it lacks estimates of fixed effects, this model cannot be used to predict the outcome of a patient treated by a different surgeon.

We note, however, that a random effects model assumes that the random effects are

[11] We follow the notation of KC and Terwiesch (2011) and use i to associate surgeon and hospital volumes with patients.

[12] 12 out of the 188 surgeons performed surgeries at multiple hospitals. Because surgeon performance is institution-specific and not fully transferable across hospitals (Huckman and Pisano 2006), we assume that these surgeons have independent unobserved effects. This assumption allows us to estimate provider quality using the multilevel probit model.

uncorrelated with other regressors in the model while a fixed effects model allows for correlation between the fixed effects and other regressors. If this assumption is wrong, estimates from the random effects model will be inconsistent and systematically different from those estimated from the fixed effects model. We check this assumption using the Hausman test (Hausman 1978) with the null hypothesis that the difference between the random effects and fixed effects estimates is not significant. A p-value of 0.575 from the Hausman test (see Appendix A.1) indicates the difference between the two sets of estimates is not statistically significant, which provides additional justification for using the random effects model.

2.4.3 Estimation of Hospital and Surgeon Effects

Because our primary goal is to help patients find better care, our main interests are the hospital specific effect α_k and the surgeon specific effect β_{jk} on patient outcomes. To estimate these effects, we use an orthogonal transformation as in Gibbons and Bock (1987). That is, we rewrite the hospital and surgeon specific effect as $\alpha_k = \mu_\alpha + \theta_k\sigma_\alpha$ and $\beta_{jk} = \mu_\beta + \theta_{jk}\sigma_\beta$, where θ_k and θ_{jk} follow the standard normal distribution.

Note that, conditional on hospital and surgeon specific effects θ_k and θ_{jk}, the outcomes of all patients treated by surgeon j at hospital k are independent; therefore, the marginal probability of observing the set of outcomes at a hospital k can be expressed as:

$$h(Y_k) = \int_{\theta_k} \Big(\prod_{j=1}^{J_k} \int_{\theta_{jk}} \Big(\prod_{i=1}^{N_{jk}} l(Y_{ijk}|X_{ijk}, \theta_{jk}, \theta_k, \mu, \sigma, \gamma) \Big) \phi(\theta_{jk}) d\theta_{jk} \Big) \phi(\theta_k) d\theta_k.$$

The individual likelihood function $l(Y_{ijk}|X_{ijk}, \theta_{jk}, \theta_k, \mu, \sigma, \gamma)$ equals:

$$l(Y_{ijk}|X_{ijk}, \theta_{jk}, \theta_k, \mu, \sigma, \gamma) = \big[\Phi(z_{ijk}(X_{ijk}, \theta_k, \theta_{jk}; \mu, \sigma, \gamma))\big]^{Y_{ijk}} \big[1 - \Phi(z_{ijk}(X_{ijk}, \theta_k, \theta_{jk}; \mu, \sigma, \gamma))\big]^{1-Y_{ijk}},$$

where

$$\begin{aligned} z_{ijk}(X_{ijk}, \theta_k, \theta_{jk}; \mu, \sigma, \gamma) &= \gamma_0 + \gamma_1 Age_i + \gamma_2 Gender_i + \gamma_3 Race_i + \gamma_4 Comorb_i, \\ &\quad + \gamma_5 SurgVol_i + \gamma_6 HospVol_i + \mu_\alpha + \theta_k\sigma_\alpha + \mu_\beta + \theta_{jk}\sigma_\beta. \end{aligned}$$

We can now estimate (γ, μ, σ) by maximizing the log likelihood of observing the outcomes at all hospitals, which is expressed as:

$$\log L = \sum_{k=1}^{K} \log h(Y_k).$$

Upon obtaining estimates $(\hat{\gamma}, \hat{\mu}, \hat{\sigma})$, we calculate $\hat{\theta}_k$ and $\hat{\theta}_{jk}$ for each hospital and each surgeon using the expected a posteriori (EAP) value (Bayes estimate) of θ_j and θ_{jk} (Bock & Aitkin 1981, Gibbon & Hedeker 1997).

$$\begin{aligned}
\hat{\theta}_k &= \frac{\int_{\theta_k} \theta_k \Big(\prod_{j=1}^{J_k} \int_{\theta_{jk}} \Big(\prod_{i=1}^{N_{jk}} l(Y_{ijk}|X_{ijk}, \theta_{jk}, \theta_k, \hat{\mu}, \hat{\sigma}, \hat{\gamma}) \Big) \phi(\theta_{jk}) d\theta_{jk} \Big) \phi(\theta_k) d\theta_k}{\int_{\theta_k} \Big(\prod_{j=1}^{J_k} \int_{\theta_{jk}} \Big(\prod_{i=1}^{N_{jk}} l(Y_{ijk}|X_{ijk}, \theta_{jk}, \theta_k, \hat{\mu}, \hat{\sigma}, \hat{\gamma}) \Big) \phi(\theta_{jk}) d\theta_{jk} \Big) \phi(\theta_k) d\theta_k}, \\
\hat{\theta}_{jk} &= \frac{\prod_{j=1}^{J_k} \int_{\theta_{jk}} \theta_{jk} \Big(\prod_{i=1}^{N_{jk}} l(Y_{ijk}|X_{ijk}, \theta_{jk}, \hat{\theta}_k, \hat{\mu}, \hat{\sigma}, \hat{\gamma}) \Big) \phi(\theta_{jk}) d\theta_{jk}}{\prod_{j=1}^{J_k} \int_{\theta_{jk}} \Big(\prod_{i=1}^{N_{jk}} l(Y_{ijk}|X_{ijk}, \theta_{jk}, \hat{\theta}_k, \hat{\mu}, \hat{\sigma}, \hat{\gamma}) \Big) \phi(\theta_{jk}) d\theta_{jk}}.
\end{aligned}$$

These quantities can be evaluated using Gauss-Hermite quadrature as described in Gibbons and Bock (1987) or Bock and Aitkin (1981). Estimates of α_k and β_{jk} can be recovered by $\hat{\alpha}_k = \hat{\mu}_\alpha + \hat{\theta}_k \hat{\sigma}_\alpha$ and $\beta_{jk} = \hat{\mu}_\beta + \hat{\theta}_{jk} \hat{\sigma}_\beta$. Finally, the standard errors can be estimated using

$$\begin{aligned}
\sigma(\hat{\theta}_k) &= \frac{\int_{\theta_k} (\theta_k - \hat{\theta}_k)^2 \Big(\prod_{j=1}^{J_k} \int_{\theta_{jk}} \Big(\prod_{i=1}^{N_{jk}} l(Y_{ijk}|X_{ijk}, \theta_{jk}, \theta_k, \hat{\mu}, \hat{\sigma}, \hat{\gamma}) \Big) \phi(\theta_{jk}) d\theta_{jk} \Big) \phi(\theta_k) d\theta_k}{\int_{\theta_k} \Big(\prod_{j=1}^{J_k} \int_{\theta_{jk}} \Big(\prod_{i=1}^{N_{jk}} l(Y_{ijk}|X_{ijk}, \theta_{jk}, \theta_k, \hat{\mu}, \hat{\sigma}, \hat{\gamma}) \Big) \phi(\theta_{jk}) d\theta_{jk} \Big) \phi(\theta_k) d\theta_k}, \\
\sigma(\hat{\theta}_{jk}) &= \frac{\prod_{j=1}^{J_k} \int_{\theta_{jk}} (\theta_{jk} - \hat{\theta}_{jk})^2 \Big(\prod_{i=1}^{N_{jk}} l(Y_{ijk}|X_{ijk}, \theta_{jk}, \hat{\theta}_k, \hat{\mu}, \hat{\sigma}, \hat{\gamma}) \Big) \phi(\theta_{jk}) d\theta_{jk}}{\prod_{j=1}^{J_k} \int_{\theta_{jk}} \Big(\prod_{i=1}^{N_{jk}} l(Y_{ijk}|X_{ijk}, \theta_{jk}, \hat{\theta}_k, \hat{\mu}, \hat{\sigma}, \hat{\gamma}) \Big) \phi(\theta_{jk}) d\theta_{jk}}.
\end{aligned}$$

2.5 Generating Patient-Specific Quality Information

In this section, we first summarize results from the above model (i.e., Equation 2.1) applied to different quality metrics. Then we examine quality gaps between surgeons and show that the quality gaps are heterogeneous for different patients. Finally, we make use of the quality-adjusted life expectancy metric to show how patients of different demographics and levels of acuity benefit differently from elite surgeons.

2.5.1 Summary Statistics and Estimation Results

Between 2009 and 2012, 2,718 patients of New York State underwent elective mitral valve surgery performed by 188 surgeons at 35 New York hospitals. Among these patients, 26% bypassed their local hospitals (i.e., those within 5 miles of the nearest hospital) and chose a hospital they had not visited for one year or more. Table 2.2 summarizes their characteristics. These data reveal some insights into patient choices. For example, the average travel distance is longer for patients under 60 than for older patients. This may be because younger patients are better able to travel. However, patients over 80 travelled on average further than patients in their 60s and 70s. This could be because their medical condition is often too delicate for a local hospital to handle. Overall, the average observed mortality rate was 1%, complication rate was 11%, readmission rate was 4%, and repair rate was 57%. However, all of these metrics worsen as the age of the patient increases.

Table 2.2: Summary of Patient Characteristics

	Patients	Travel Dist. (miles)		Repair Rate		Mortality Rate		Complication Rate		Readmission Rate	
	%	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.
Age											
below 50	12%	19	29	67%	47%	0%	5%	5%	22%	3%	17%
50 to 60	22%	22	29	72%	45%	1%	7%	5%	22%	2%	15%
60 to 70	27%	18	26	60%	49%	2%	13%	9%	29%	4%	21%
70 to 80	27%	17	22	47%	50%	1%	12%	15%	36%	5%	21%
above 80	13%	19	22	39%	49%	3%	16%	20%	40%	9%	28%
Gender											
male	55%	20	26	63%	48%	1%	11%	11%	31%	4%	20%
female	45%	18	27	51%	50%	1%	12%	11%	31%	5%	22%
Race											
asian	2%	18	22	42%	50%	2%	14%	9%	30%	8%	27%
black	8%	8	19	52%	50%	1%	12%	16%	37%	4%	19%
hispanic	5%	8	17	44%	50%	0%	0%	11%	32%	5%	22%
others	13%	18	22	64%	48%	1%	12%	10%	31%	10%	30%
white	73%	21	28	58%	49%	1%	11%	10%	30%	3%	18%
Total	2,718	19	26	57%	49%	1%	11%	11%	31%	4%	21%

To compute risk adjusted quality metrics, we estimate the quality model (i.e., Equation 2.1) from Section 2.4. Table 2.3 summarizes the results. We first examine how patient characteristics affect outcomes. Not surprisingly, repair rate decreases and rates of mortality, complication and readmission increase as patient age increases. Compared with male patients, female patients are less likely to receive mitral valve repair, and are more likely to have deaths. Compared with white patients, Hispanic patients are less likely to receive mitral valve repair, and are more likely to have complications. Finally, mitral valve repair rate is lower for patients with comorbidities of atrial fibrillation, chronic lung disease or renal disease. These results are consistent with those of previous studies (see e.g., Bolling et al., 2010 and Vassileva et al. 2013). Comorbidities such as atrial fibrillation and chronic lung disease affect other measures of quality as well, but the impact and significance level vary for different quality metrics.

Surgical volume also influences outcomes. For mitral valve repair, we see that the coefficients of both hospital and surgeon volumes are positive and significant at the 5% level, suggesting that repair rate increases with volume. For complication and readmission, we see that complication rate decreases with surgeon volume and readmission rate decreases with hospital volume. These results are consistent with those of previous studies (see e.g., Birkmeyer et al. 2003 and Kilic et al. 2013). For mortality, we do not observe a statistically significant effect of volume, partly because the events measured are relatively rare. In addition, we do see that repair leads to a lower level of complication than replacement, as suggested by the medical literature (LaPar et al. 2010).[13]

Table 2.3: Estimation Results of the Quality Model

	Repair		Mortality		Complication		Readmission	
	Coefficient	Std. Err.	Coefficient	Std. Err.	Coefficient	Std. Err.	Coefficient	Std. Err.
Surgical Volumes								
hosp_vol	0.21 ∗∗	0.09	0.14	0.12	−0.05	0.04	−0.18 ∗∗	0.08
surg_vol	0.13 ∗∗∗	0.04	−0.01	0.04	−0.04 ∗∗	0.02	0.00	0.05
Patient Demographics								
age	−0.02 ∗∗∗	0.00	0.02 ∗∗∗	0.01	0.02 ∗∗∗	0.00	0.02 ∗∗∗	0.00
female	−0.31 ∗∗∗	0.07	0.24 ∗∗	0.10	0.06	0.05	0.12	0.10
black	−0.17	0.17	0.14	0.19	0.14	0.15	0.14	0.17
hispanic	−0.32∗	0.17	−5.45		−0.01	0.13	0.29	0.20
asian	−0.29 ∗∗	0.15	0.07	0.47	−0.07	0.17	0.62∗	0.34
others	0.13	0.09	−0.09	0.22	−0.21 ∗∗	0.10	0.62 ∗∗∗	0.25
Comorbidities								
atrial fibrillation	−0.12∗	0.07	−0.57 ∗∗∗	0.12	0.01	0.12	0.03	0.07
heart failure	−0.09	0.34	0.53	0.61	1.70 ∗∗	0.74	−4.30 ∗∗∗	0.18
lung disease	−0.23 ∗∗∗	0.08	−0.02	0.27	−0.01	0.08	0.28 ∗∗∗	0.09
diabetes	−0.12	0.08	−0.37	0.24	0.11	0.08	−0.17∗	0.09
hypertension	0.05	0.05	−0.43 ∗∗∗	0.15	−0.31 ∗∗∗	0.08	0.02	0.09
renal failure	−0.29 ∗∗∗	0.08	0.38	0.26	0.89 ∗∗∗	0.11	0.23	0.15
Others								
repair			−0.26	0.21	−0.25 ∗∗∗	0.08	0.11	0.12
θ_α	0.14	0.03	0.00	0.00	0.05	0.03	0.04	0.04
θ_β	0.10	0.03	0.00	0.00	0.00	0.00	0.06	0.10
constant	0.79 ∗∗∗	0.31	−4.24 ∗∗∗	0.77	−2.38 ∗∗∗	0.28	−2.54 ∗∗∗	0.34
log likelihood	−1544.29		−127.78		−855.70		−445.89	

Note: ∗∗∗ $p < 0.01$, ∗∗ $p < 0.05$, ∗ $p < 0.1$. Robust standard errors are clustered by surgeon. The following comorbidities are included in the regression but are not shown in the table: alcohol abuse, deficiency anemias, rheumatoid arthritis/collagen vascular diseases, chronic blood loss anemia, coagulopathy, depression, drug abuse, hypothyroidism, liver disease, lymphoma, fluid and electrolyte disorders, metastatic cancer, other neurological disorders, obesity, paralysis, peripheral vascular disorders, psychoses, pulmonary circulation disorders, solid tumor without metastasis, valvular disease, and weight loss.

2.5.2 Population-Average vs. Patient-Specific Outcomes

We estimate population-average outcomes of surgeons using the predicted rates of mitral valve repair, complication, readmission and mortality for a patient with average

[13] We checked the quality model for these three measures without including repair as an explanatory variable, and obtained similar quality gaps between surgeons.

characteristics. We follow the approach of New York Cardiac Surgeon Report Card to determine whether a surgeon is statistically significantly different from the state average, which is defined as the mean of all surgeons rates for that measure.[14] We informally label the top performers, who are significantly better than the state average, as "elite surgeons".[15] Figure 2.1 displays mitral valve repair rate as an example (readmission and complication rates are shown in Appendix A.2). The average repair rate across all surgeons is around 50%. The confidence intervals are heavily influenced by the number of cases. Surgeons with low volumes tend to have wide confidence intervals, and are therefore either indistinguishable from the state average or below it. While almost all elite surgeons are high-volume surgeons, not all high-volume surgeons have high mitral valve repair rates.

Figure 2.1: Mitral Valve Repair Rate by Surgeon for A Patient with Average Characteristics

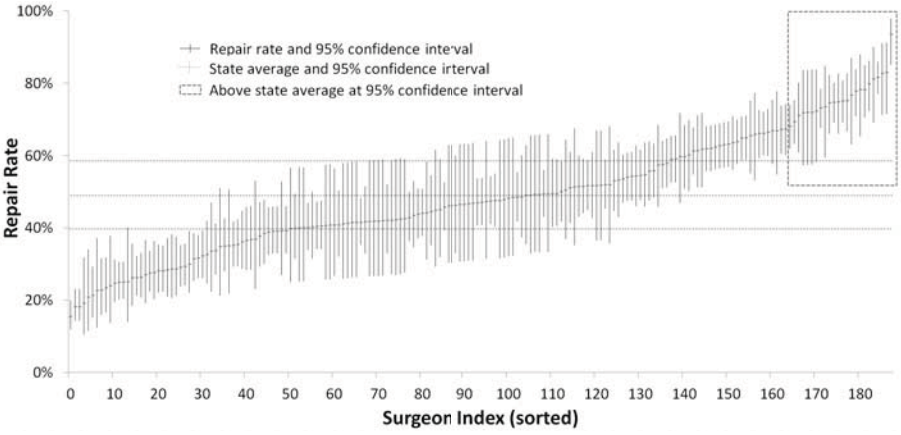


Figure 2.1, as well as the figures in Appendix A.2, clearly suggest that some surgeons are significantly better than others. What is not obvious is that the quality gap between elite and other surgeons is not uniform across patients of different demographics and levels of acuity. To illustrate this point, we define three patient types with different levels of acuity: "sick" (i.e., 90 years old with comorbidities) patients, "typical" (i.e., average) patients, and "healthy" (i.e., 30 years old with no comorbidities) patients. We calculate the predicted rates of mitral valve repair for the sick and healthy patients in

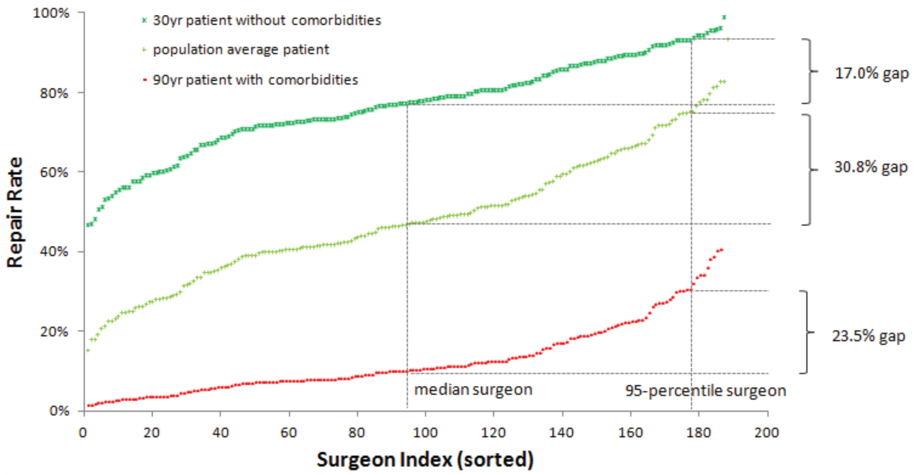
[14] An alternative way to calculate the state average is to weight an individual surgeon's rate by his/her surgical volume. However, because surgeons are the focus of this analysis and surgical volume is endogenous and can change over time, we calculate the averages at the surgeon rather than the patient level.

[15] We note that there are alternative ways to define elite surgeons (e.g., using a cut-off rate). These alternative definitions do not change our main conclusion regarding the value of patient-specific information.

the same way as what we did for the typical patients. Figure 2 shows mitral valve repair rate as an example (complication and readmission rates are shown in Appendix A.3). In this case, all three groups of patients benefit from visiting elite surgeons, but the magnitude of benefit differs. For example, the gap in repair rate between a surgeon at the 95th-percentile and a surgeon at the median for these three patient types is 30.8% (p=0.032) for the typical patients, 17.0% (p=0.098) for the healthy patients and 23.5% (p=0.099) for the sick patients, all of which are significant at 10% significance levels.[16]

This indicates that neither the sick nor the healthy group of patients benefit as much as the typical patients from visiting an elite surgeon. A plausible explanation for this is that many sick patients, with hard-to-repair valves, are likely to get a replacement regardless of which surgeon they visit, while many healthy patients, with easy-to-repair valves, are likely to get a replacement from any above-median surgeon. The patients in between, however, tend to present "difficult but not impossible" repair challenges, and are therefore substantially more likely to receive a repair from an elite surgeon than from a median surgeon. Of course, these three sample groups are only illustrative. The heterogeneity in surgeon impact on patient outcomes may be affected by many patient characteristics.

Figure 2.2: Mitral Valve Repair Rate by Surgeon for Patients of Different Levels of Acuity



[16] The quality gap between the top (100th-percentile) surgeon and a surgeon at the median is also heterogeneous with the healthy patients benefit less than the sick and typical patients. Our simulation model to be discussed later captures the quality gap between any two surgeons across all patient types.

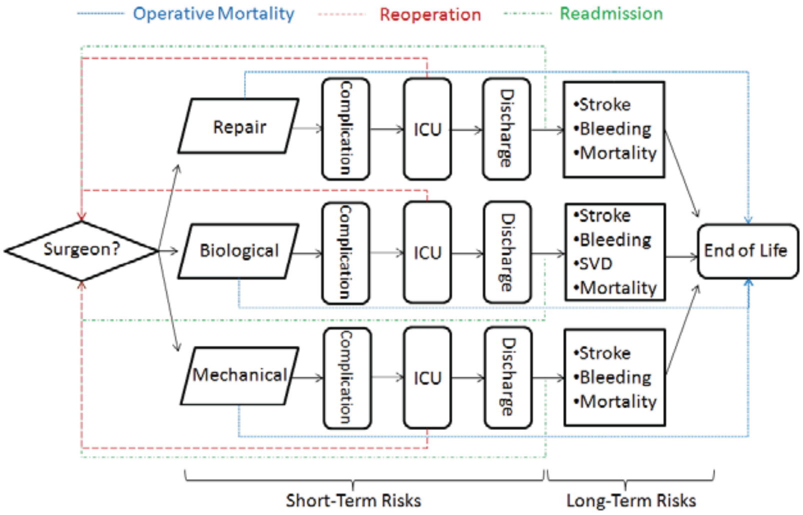
2.5.3 Quality-Adjusted Life Expectancy

The single number quality metrics we have used so far are common in the medical literature (mortality, complication and readmission rates for all surgical procedures and repair rate for mitral valve surgery). Examining how these vary by surgeon and patient type highlights the heterogeneity we must quantify to generate patient-specific outcome information. But none are entirely satisfactory on their own as characterizations of the patient experience. One reason is that mortality, complication and readmission rates capture only short-term issues. Repair rate captures long-term quality of life, but only partially. It omits the intensity and duration of the quality of life benefits of a repair. For example, younger and healthier patients have more years to live and, as a result, their lifetime benefit from a successful repair is greater than that of older and sicker patients. For patients with the same number of years to live, those suffering from complications such as stroke or bleeding have a lower quality of life than those without such complications. To account for such differences, we make use of the Quality-Adjusted Life Expectancy (QALE) metric (see e.g., Black et al. 2014, Hutton et al. 2011, Zaric et al. 2000).

To compute QALE, we make use of the model in Figure 2.3 to capture both short- and long-term postoperative risks associated with mitral valve surgery. Short-term risks include operative mortality, complications (e.g., stroke, wound infection, renal failure, reoperation and ventilation) observed during hospitalization, and 30-day readmission. Long-term risks include stroke, bleeding, reoperation due to valve deterioration, and mortality incurred during the remainder of a patient's life. We assess these risks of treatment by elite and other surgeons according to patients' age and major comorbidities (e.g., atrial fibrillation, heart failure, lung disease, diabetes, hypertension and renal failure).

To calculate the quality-adjusted life expectancy of patient type i treated by surgeon j (denoted as $QALE_{ij}$), we let $QoL_{ij}(t)$ denote the patient's quality of life at time $t, t \in [0, T_{ij}]$, where 0 is the time of treatment, and T_{ij} is the survival time. Then the patient's quality-adjusted life expectancy can be described as $QALE_{ij} = \int_0^{T_{ij}} QoL_{ij}(t)dt$. The right-hand side of this equation cannot be calculated directly, because the upper bound of integration T_{ij} depends on patient survival. To address this issue, we re-write $QALE_{ij}$ as a function of the survival function $S_{ij}(t)$. Let TR_{ij} denote the treatment (repair or replacement) received by patient i at surgeon j. $R_{ij}(t, TR_{ij})$ denote the occurrence of one or more major risks to patient i at time t after being treated by surgeon j, where the set of major risks include mortality, readmission, and complications. In particular, we define R_{ij}^s and R_{ij}^l to distinguish short-term and long-term risks

Figure 2.3: Related Events and Decision Process for Patients with Mitral Valve Diseases



Note: ICU and SVD stand for intensive care unit and structural valve deterioration, respectively.

described earlier. Let $t \leq t_1$ and $t > t_1$ indicate short and long terms, respectively.

We then calculate the expected $QALE_{ij}$ as follows (Hwang et al. 1996)

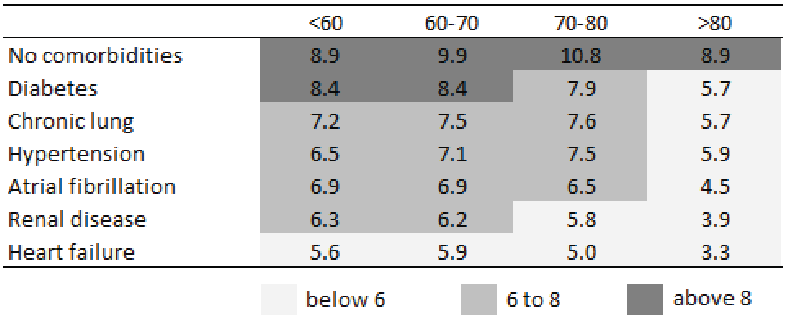
$$\begin{aligned} QALE_{ij} &= \int_0^\infty S_{ij}(t)QoL_{ij}(t)dt \qquad (2.2) \\ &= E_{TR_{ij}}\left[\int_0^\infty S_{ij}(t|TR_{ij})E_{R_{ij}(t,TR_{ij})}\left[QoL_{ij}(t|R_{ij}(t,TR_{ij}))\right]dt\right] \\ &= E_{TR_{ij}}\left[\int_0^{t_1} S_{ij}(t|TR_{ij})E_{R_{ij}(t,TR_{ij})}\left[QoL_{ij}(t|R^s_{ij}(t,TR_{ij}))\right]dt\right. \\ &+ \left.\int_{t_1}^\infty S_{ij}(t|TR_{ij})E_{R_{ij}(t,TR_{ij})}\left[QoL_{ij}(t|R^l_{ij}(t,TR_{ij}))\right]dt\right] \end{aligned}$$

To parameterize this model, we estimate operative mortality, short-term complication and 30-day readmission rates using quality models similar to Equation 2.1 presented in Section 2.4.3. In particular, we estimate the probability of each type of complication separately and consider the occurrence of each complication as independent when calculating the joint probability of multiple complications (see Appendix A.4). We also distinguish biological replacement from mechanical replacement in calculating the probability of receiving such treatment, as well as the state of health conditional on each type of treatment.

Because the current data does not allow us to analyze long-term risks and quality of life, we follow approaches of existing studies (see e.g., Black et al. 2014 and Zaric et al. 2000) to estimate them from several sources in the medical literature. We estimate risks of stroke, bleeding, structural valve deterioration and mortality based on Bourguignon et al. (2014), Daneshmand et al. (2010), Gelsomino et al. (2011), Ray et al. (2006), Ruel et al. (2004), Russo et al. (2008). We estimate quality of life based on Cox et al. (2007), Jideus et al. (2009), Regier et al. (2006), Shah and Gage (2011), Sullivan and Ghushchyan (2006), Windisch et al. (2003). Details about these resources and value of each model element are provided in Appendix A.5.

To illustrate the heterogeneous benefits patients obtain from elite surgeons, Figure 2.4 summarizes the quality-adjusted life expectancy a patient gains from visiting a 95th-percentile instead of 50th-percentile surgeon. In this case, the additional benefit from visiting the elite surgeon ranges from 3.3–10.8 months depending on patients' age and comorbidities. Generally speaking, younger and healthier patients benefit more, because they have more years to live and their quality of life is higher. However, the relationship between patients' age and benefits is not monotonic, because neither the quality gap between surgeons nor quality of life difference between procedures is a monotonic function of age.[17] Of course, these are only illustrative examples. The expected gain in quality-adjusted life years will differ across all patient groups and different surgeons. The model (i.e., Equation 2.1) described in Section 4 can be used to estimate surgeon performance, and hence quality gaps, for all patient groups.

Figure 2.4: Difference in Quality-Adjusted Life Expectancy (Months) between 95th-percentile and 50th-percentile Surgeons



	<60	60-70	70-80	>80
No comorbidities	8.9	9.9	10.8	8.9
Diabetes	8.4	8.4	7.9	5.7
Chronic lung	7.2	7.5	7.6	5.7
Hypertension	6.5	7.1	7.5	5.9
Atrial fibrillation	6.9	6.9	6.5	4.5
Renal disease	6.3	6.2	5.8	3.9
Heart failure	5.6	5.9	5.0	3.3

□ below 6 ▨ 6 to 8 ■ above 8

Note: These results will be used to describe patient utility in the next section when we formulate patients' choice of surgeon as a queueing system. The relationship between patients' age and benefits is not monotonic, because neither the quality gap between surgeons nor quality of life difference between procedures is a monotonic functions of age.

[17] One reason for this is that if an older patient must receive a replacement valve, he/she will receive a biological implant. A younger patient will receive a mechanical implant, which will last longer but carries risk of clots causing a stroke.

2.6 Estimating the Value of Patient-Specific Information

The primary goal of a rating system that scores or ranks health care providers is to help ensure that patients are treated by an appropriate provider. It is almost tautological that better information about provider quality should enable better matching of patient to provider. How much it helps, however, depends on how capacity constraints on the elite providers affect the allocation of their services to patients. Hence, we must first model this allocation before we can evaluate the value of patient centric information in provider rankings.

Capacity allocation is straightforward in centrally planned healthcare systems, such as the nationalized UK system or the single-payer Canadian system, where priorities can be set according to various criteria (e.g., age, health status) in the pursuit of a socially optimal allocation. Under such a system, patient-specific quality information simply provides another criterion (i.e., medical benefit from treatment by an elite provider) that can be used in the optimization. But we leave analysis of the impact of patient-specific information on centrally planned healthcare systems to others and focus instead on the US.

The hybrid healthcare system of the US occasionally uses central planning (e.g., to set priorities for transplant organs), but usually relies on market mechanisms, under which patients pay (e.g., for elective cosmetic surgery) or wait (e.g., to get on the schedule of a busy surgeon). In the case of mitral valve disease, there is no centrally planned system for assigning patients to surgeons. Also, because (unlike elective cosmetic surgery) mitral valve surgery is covered by medical insurance, surgeons do not compete on price. Hence, the allocation of elite mitral valve surgeon time relies on patient waiting time. It will generally take longer to be treated by a busy elite surgeon than by a less busy average surgeon. To account for this, we will make use of a choice model in which patients consider surgeon quality and waiting time, as well as travel distance, in selecting a surgeon.

From a policy standpoint, we are particularly interested in the effect of patient-specific information on how elite surgeon capacity is allocated and the impact this has on overall patient outcomes. Rankings based on population-average information imply that the clinical benefit from treatment by an elite surgeon is the same for all patients. Hence, who winds up being treated by an elite surgeon will be determined only by travel distance and waiting time. Patients who can wait longer or travel further will be more likely to go to an elite surgeon, irrespective of how much they benefit from doing so.

In contrast, rankings based on patient-specific information reflect the differences in the clinical benefit from treatment by an elite surgeon across patient types. Therefore, patients who make use of such rankings will be able to consider their personal clinical benefit, as well as waiting time and travel distance, in choosing a surgeon. So, if waiting time and travel distance are equal for two patients, the patient who benefits more from treatment by an elite surgeon is more likely to wait for and travel to that elite surgeon. Note that no coercion or optimization is needed to achieve the improvement in social outcome. Patients make their own choices to maximize personal utility. But, since patients who benefit more from an elite surgeon will wait longer and travel further, the elite surgeons will naturally wind up treating the patients who benefit most from their specialized skills.

2.6.1 Patient Choice Model

To evaluate the magnitude of the overall benefit to society of patient-specific quality information in mitral valve surgery, we make use of a patient choice model, in which patients select providers based on quality (QALE), distance, and wait time. To model wait time, we represent surgeons and patients as a system of parallel M/M/1 queues. We define a server as a cardiac surgeon who performs mitral valve surgery, so there are in total 188 servers in this study. The service rate of surgeon j, denoted as μ_j, is defined as the maximum number of elective mitral valve surgeries he/she can perform in a month. The arrival rate of patients, denoted as λ, is defined as the monthly rate at which patients in need of elective mitral valve surgery arrive at one of the 188 surgeons. Patient i decides whether to join queue j based on the expected quality-adjusted life expectancy from surgeon j ($QALE_{ij}$), the travel distance to surgeon j ($Distance_{ij}$) and the waiting time ($WaitTime_{ij}$). We compute the expected quality-adjusted life expectancy with population-average and patient-specific information so we can make comparisons. Assuming service times are exponential, we can express patient i's expected waiting time for surgeon j as $(N_{ij}+1)/\mu_j$, where N_{ij} represents the number of patients in queue ahead of patient i. Patient i chooses a surgeon that maximizes his/her utility:

$$\begin{aligned} \arg\min_j Utility_{ij} &= QALE_{ij} - \alpha Distance_{ij} - \beta WaitTime_j \qquad (2.3)\\ &= QALE_{ij} - \alpha Distance_{ij} - \beta(N_{ij}+1)/\mu_j \end{aligned}$$

Obviously this model simplifies reality in many ways. Most importantly, it assumes the relative weights patients place on quality-adjusted life expectancy, distance and waiting time are the same across all patients. In reality, patient preferences may vary depending on age, medical condition and other individual characteristics. However, as we will show later, considering heterogeneous patient preferences only reinforces our main conclusions.

For the case of population-average information, we let $\overline{Rate}_j$ denote the population-average repair rate for provider j and compute $QALE_{ij} = q_i\overline{Rate}_j$, where q_i indicates expected increase in quality-adjusted life expectancy for patient i per percentage point increase in repair rate. A patient chooses a surgeon under the assumption that his/her utility is computed using $q_i\overline{Rate}_j - \alpha Distance_{ij} - \beta(N_{ij}+1)/\mu_j$. For the case of patient-specific information, we let $Rate_{ij}$ denote the likelihood of a repair when patient i is treated by surgeon j. A patient chooses a surgeon under the assumption that his/her utility is computed using $q_iRate_{ij} - \alpha Distance_{ij} - \beta(N_{ij} + 1)/\mu_j$.

To evaluate the impact of patient-specific information at the social level, we define the social value as the total patient utility (i.e., $SocialValue = \sum_i[\max_j(QALE_{ij} - \alpha Distance_{ij} - \beta WaitTime_j)]$). Then we simulate the queueing system based on parameter values that are consistent with empirical data. As described previously, the service rate of a surgeon is defined as the maximum number of elective mitral valve surgeries he/she can perform in a month as observed from the data. The arrival rate of patients is defined as the average number of patients treated by any of the 188 surgeons in a month. The units of $QALE_{ij}$, $Distance_{ij}$ and $WaitTime_j$ are days, miles and months, respectively. Specifically, the quality-adjusted life expectancy for patients at different ages and with different comorbidities is obtained from the model (i.e., Equation 2.2) discussed earlier.

Because the weight on quality-adjusted life expectancy has been normalized to one, the values of α and β can be interpreted as equivalent quality-adjusted life days per mile and month, respectively. We reviewed the relevant literature to calibrate reasonable values of α and β. Finlayson et al. (1999) reported that most patients are willing to travel to a regional hospital if this will reduce mortality rate by 20%. The baseline mortality rate and travel distance were not specified. But we can make a rough estimation of the α coefficient if we assume that the mortality rate at a local hospital is 2%, the extra travel distance to a regional hospital is 30 miles, and a patient's remaining life expectancy is 20 years. Under these assumptions, the increase in life days from a 20% reduction of the 2% mortality rate is 20 yrs × 365 days/yr × 0.4% = 29.2 days. This suggests that α is around 29.2 life days ÷ 30 miles = 0.97 life days per mile. To create a range around this estimate, we specify three levels of weights on travel distance (α), denoted as Low, Medium and High, to be 0.5, 1 and 5. Similarly, in another study, Dixon et al. (2010) reported that patients value each month of waiting time as worth 40–60 minutes of travel time. This implies the weight patients place on waiting time is 40–60 if we assume that 1 hour of travel is roughly equivalent to 60 miles of distance. Again, to create a plausible range, we specify three levels of weights on waiting time (β), denoted as Low, Medium and High, to be 10, 50 and 100. Other results that support these estimates can be found in Burge et al. (2005) and Gaynor et al. (2010).

To simulate the arrival of patients, we randomly draw with replacement from the 2,718 patients in this study. The draws occur according to a Poisson process with an average rate of 57 (i.e., 2,718/48) per month. To evaluate queueing in steady state, we skip the first 1,000 observations (i.e., the warm-up period) and focus on the subsequent 5,000 patients for analyses.

2.6.2 Simulation Results

We consider 27 different variants of the simulation described earlier (three capacity levels, three distance weights and three waiting time weights) under both population-average and patient-specific information. The results are summarized in Table 2.4, where columns (1) and (2) show the weights on travel distance and waiting time, columns (3)–(7) show the expected number of repairs, and columns (4)–(8), (5)–(9) show the average travel distance and waiting time per patient. To compare results across different scenarios, we convert utilities to Convenience Adjusted QALEs, which represent equivalent quality-adjusted life days after adjusting for inconvenience of travelling and waiting. The results are summarized in (6)–(10).

To interpret these results, we first recall that the 2,718 New York mitral valve patients we considered in our empirical analysis traveled an average distance of 19 miles to receive surgery and that 1,557 (or 57%) had their valves successfully repaired. In terms of clinical outcomes this is worse than the case with population-average information and High weights on distance and waiting time (because equivalent quality-adjusted life days per mile is 5, and equivalent quality-adjusted life days per month is 100), which results in 2,057 (or 76%) repairs, an average travel distance of 18 miles and an average waiting time of 15 days.

At first glance, this might suggest that patients behaved as if they were strongly travel and wait averse. But it doesn't seem reasonable that an average patient would be willing to sacrifice more than 5 days of life to avoid traveling 1 extra mile and

Table 2.4: Comparison of the Effectiveness of Patient-Specific Information and Capacity Increase

Weight on Distance (1)	Weight on Waiting Time (2)	Expected Number of Repairs (3)	Average Distance (miles) (4)	Average Waiting (months) (5)	Convenience Adjusted QALE*(days) (6)	Expected Number of Repairs (7)	Average Distance (miles) (8)	Average Waiting (months) (9)	Convenience Adjusted QALE*(days) (10)
		Patient-Specific (current capacity)				Population-Average (current capacity)			
	Low	2,174	26	5.8	258	2,104	26	5.7	244
Low	Medium	2,160	28	1.1	255	2,101	28	1.1	242
	High	2,152	30	0.5	256	2,095	30	0.5	244
	Low	2,162	22	5.6	247	2,103	22	5.8	234
Medium	Medium	2,151	23	1.1	244	2,095	24	1.1	232
	High	2,147	24	0.5	246	2,095	25	0.5	234
	Low	2,135	17	5.3	178	2,076	17	5.5	164
High	Medium	2,118	17	1.0	174	2,060	17	1.1	160
	High	2,099	17	0.5	173	2,057	18	0.5	162
		Population-Average (10% capacity increase)				Population-Average (20% capacity increase)			
	Low	2,127	26	5.7	252	2,142	26	5.6	256
Low	Medium	2,122	28	1.1	251	2,141	28	1.1	257
	High	2,121	30	0.5	251	2,135	30	0.5	257
	Low	2,120	22	5.4	242	2,141	23	5.5	247
Medium	Medium	2,114	24	1.1	240	2,131	24	1.1	245
	High	2,110	25	0.5	241	2,132	25	0.5	248
	Low	2,094	17	4.8	175	2,117	18	4.7	182
High	Medium	2,086	18	1.0	169	2,102	18	1.0	175
	High	2,077	18	0.5	168	2,104	18	0.5	176
Actual		1,557	19						

Note: This table compares scenarios when patients choose surgeons based on patient-specific information (with current capacity) and population-average information (with 0–20% capacity increases). We consider Low, Medium and High weights patients place on travelling and waiting. Equivalent quality-adjusted life days per mile for Low, Medium and High weights on travelling are 0.5, 1, 5. Equivalent quality-adjusted life days per month for Low, Medium and High weights on waiting are 10, 50, 100. *For the ease of comparison, a fixed amount of one quality-adjusted life years has been subtracted from Convenience Adjusted QALE for both patient-specific and population-average cases.

waiting 1.5 more days. Indeed, many studies of patient choices of health care providers have found that patients are willing to travel and wait for better care. For example, Finlayson et al. (1999) found that most patients are willing to bypass a local hospital in favor of a more distant regional hospital if this will result a 20% reduction in the likelihood of mortality. Groux et al. (2014) found that 36%–41% of cancer patients surveyed are willing to travel any distance to receive the best available treatment. Gaynor et al. (2010) found that the average waiting time for coronary artery bypass surgery is around 2–3 months. With these in mind, we conclude that the behavior of New York patients is not primarily driven by travel or wait aversion.

A second possible explanation for the failure of patients to travel or wait for better care is that their choices are limited by their insurance providers. But our empirical analysis showed that patients with non-restrictive coverage (e.g., Medicare) are almost as likely to receive inferior local treatment as are patients with more restrictive (e.g., HMO) coverage.[18] So, while insurance may play a role, it does not seem to be the dominant driver of surgeon choice.

A third explanation is that patients either fail to find outcome data, or, if they do, fail to understand or trust it. Without information with which to distinguish the performance of different surgeons, patients fall back on other criteria like convenience or

[18] Based on NY data from 2009–2012, we calculated that 64% of Medicare and 63% of HMO patients were treated by non-elite surgeons.

familiarity when choosing a surgeon. Since the studies cited above imply that patients are willing to act on quality information if they have it, many scholars (e.g., Emmert and Schlesinger 2016, Sinaiko et al. 2012) have conjectured that better presentation of medical quality information would cause patients to make more use of it in their decisions. A feature that is often cited as desirable is personalized information tailored to individual patients (Paddock et al. 2015, Sinaiko et al. 2012). The implication is that more usable information could make the better outcomes in Table 2.4 (i.e., those with low weights on distance and waiting time) possible.

Table 2.5 compares the values of using population-average and patient-specific information under current surgeon capacity. Columns (1) shows the weights on distance. Columns (2)–(4) show the change in the number of mitral valve repairs for Low, Medium and High weights on waiting time, respectively. Columns (5)–(7), (8)–(10) and (11)–(13) show the changes in the number of total quality-adjusted life years, average travel distance and waiting time per patient when information is switched from population-average to patient-specific for Low, Medium and High weights on waiting time.

To interpret these results, we use Medium weights on travel distance and waiting time for illustration. In this scenario, the expected number of repairs under population-average information is 2,095 (see Table 2.4). To achieve this, patients would have to travel an average of 24 miles and wait around 1.1 months. If patient-specific information is used, then the expected number of repair increases to 2,151 (a 2.7% increase). The additional 56 repairs relative to the population-average information case is the result of more patients who benefit most being treated by elite surgeons. Importantly, achieving this better outcome does not require additional travel distance or waiting time. It simply requires using different math to compute provider rankings.

From Table 2.5, we see that, depending on the weights patients place on travel distance and waiting time, patient-specific information increases the number of repairs by 42–70, increases total quality-adjusted life years by 56–105, changes average travel distance per patient by less than 1 mile, and reduces average waiting time per patient by 0–0.2 months. Intuitively, the overall benefits of patient-specific information tend to increase as the weights on distance or waiting time decrease, since patients become more willing to wait and travel for higher quality care.

Another way to assess the value of patient-specific information is to look at how much physician capacity must increase under population-average information to achieve the same results with patient-specific information and current capacity. To compute this, we increase service rates under population-average information and re-simulate

Table 2.5: Comparison of the Values from Using Patient-Specific and Population-Average Information

	Diff. in Num. of Repairs			**Diff. in Total QALE (year)**			**Diff. in Average Travel Dist.(mile)**			**Diff. in Average Wait Time(month)**		
Weight on	Weight on Waiting			Weight on Waiting			Weight on Waiting			Weight on Waiting		
Distance	Low	Medium	High	Low	Medium	High	Low	Medium	High	Low	Medium	High
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
Low	70	60	57	105	88	87	0	0	1	0.0	0.0	0.0
Medium	59	56	52	81	79	74	0	0	0	-0.2	0.0	0.0
High	60	58	42	86	88	56	0	0	0	-0.2	0.0	0.0

Note: This table summarizes the changes in total number of repairs, quality-adjusted life years, average travel distance and waiting time per patient when information is switched from population-average to patient-specific. We consider Low, Medium and High weights patients place on travelling and waiting. Equivalent quality-adjusted life days per mile for Low, Medium and High weights on travelling are 0.5, 1, 5. Equivalent quality-adjusted life days per month for Low, Medium and High weights on waiting are 10, 50, 100.

the queues. We find, for instance, with Medium weights on distance and waiting time, a 20% increase in surgeon capacity results in an average convenience adjusted QALE of 245 days. This is almost identical to the results with current capacity under patient-specific information. In the population-average case, the improvement is achieved by enabling the elite surgeons with the highest repair rates to treat more patients. In the patient-specific case, the improvement is achieved by having the elite surgeons treat the right patients. From Table 2.4, we see that the average convenience adjusted QALEs from using patient-specific information are comparable to those achievable by enabling the best surgeons to treat 10%–20% more patients under population-average information, depending on the weights patients place on distance and waiting time. While increasing surgeon capacity is likely to be expensive (e.g., involving adding operating rooms or surgical staff personnel) or impossible, providing patient-specific information is merely a matter of math.

As we noted earlier, patients in the real world will place different weights on travel distance and waiting time. Dixon et al. (2010) found that old patients are more averse to travelling than young patients and equivalent travel distance per month of waiting are 40 miles for patients above 60 years old and 56–80 miles for patients below 60 years old. To account for these heterogeneous preferences, we define the disutility to patients above 60 as $Disutility_{ij}^{old} = AdverseEvent_{ij} + 1.25\alpha Distance_{ij} + \beta WaitTime_j$ and that to patients below 60 as $Disutility_{ij}^{young} = AdverseEvent_{ij} + 0.75\alpha Distance_{ij} + \beta WaitTime_j$. This means that patients below 60 are likely to travel 67% further than those above 60. The simulation results for different values of α and β are summarized in Appendix A.6. We see that the relative gap between convenience adjusted QALEs achievable from using patient-specific and population-average information and the percentage increase in capacity needed for population-average information to achieve the same benefit as patient-specific information are similar to the results in Table 2.5. We obtain consistent results when we vary the weights of waiting time similarly. If anything, heterogeneous weights on travel distance and waiting time make it easier to distribute patients among surgeons in a patient pleasing manner.

2.6.3 Generalizing Our Results

Although we have restricted our analysis in this paper to elective mitral surgery in New York State, our methodology is applicable to a wide range of health care settings. The key prerequisite for applying our approach to a given medical procedure is an outcome metric that (a) accurately represents quality from a patient perspective and (b) is measured and recorded consistently across providers at the individual patient level. In this paper, we used quality adjusted life expectancy (QALE) because it takes into account short-term experience (e.g., complications), long-term survival (life expectancy), and functional outcome (repair or replacement). We were able to compute QALE by combining statistics on mortality, readmission, complication and repair rate with results from the literature that characterize the impact of surgical outcomes on life expectancy and medical complications. While not trivial, similar approaches could be used to evaluate QALE for other surgical procedures.

However, QALE is not the only reasonable option for an outcome metric. It is plausible to use either a less or more sophisticated metric. The less sophisticated category includes the commonly used metrics of mortality, readmission and complication rates. While none of these is a comprehensive metric, they may be useful proxies in certain settings. In procedures where mortality rates are high (e.g., laparotomy, partial colectomy, liver transplants), mortality rate may be a reasonable proxy for quality. In procedure where mortality rates are very low and functional outcomes are not highly variable (e.g., coronary artery bypass graft, cataract surgery, inguinal hernia repair), complication rate may serve as a proxy for quality. Since mortality and complication rates are universally tracked, it would be straightforward for hospital rating sites, such as those maintained by the Centers for Medicare and Medicaid Services, Leapfrog Group, the Society of Thoracic Surgeons and Consumer Reports, to compute and report patient specific versions of these traditional statistics.

Our approach could also be used to generate even more patient-specific outcome statistics by allowing individual patients to choose customized weights. A simple version of this could be achieved by using a weighted average of mortality rate, complication rate, and functional outcome as the outcome metric. For example, two patients considering total hip replacement surgery might choose very different weights based on their risk preferences. A patient primarily concerned about mobility will place a heavy weight on functional outcome (measured by the Harris Hip Score or Oxford Hip Score, see Nilsdotter and Bremander 2011), while a patient worried about surgical side effects will shift weight from functional outcome to complication rate. Customized weights could also be used within a QALE metric to adjust weights of the short- and long-term factors.

2.7 Conclusion and Managerial Implications

The past decade has seen increasing efforts by the US government, payers and health care providers to improve health care quality and control health care costs. Significant energy has been devoted to improving information transparency in the hope of helping patients find the best providers. In addition to being complex and difficult to use, currently available quality information about healthcare providers is based on population averages. Our results show that population-average information is valuable to patients, but that patient-specific quality information is even more valuable. Used properly, patient-specific information can help patients find the providers that are best for them and for society as a whole.

This study addresses the challenges of measuring provider quality from a patient-specific perspective and use it to help patients find better care. With mitral valve surgery as the clinical setting, we studied the quality of cardiac surgeons in New York based on different quality metrics, including a new quality-adjusted life expectancy (QALE) metric that incorporates both short- and long-term effects. We used a multilevel probit model to capture hospital and surgeon volume effects, as well as their specific effects, on patient outcomes. This analysis shows that some surgeons are performing statistically significantly better than the state average, but that patients of different demographics and levels of acuity benefit differently from these elite surgeons.

We compared the effectiveness of providing patient-specific quality information and that of increasing surgeon capacity when patients choose a surgeon that maximize their utility. We estimate that providing patient-specific quality information in place of population-average information offers societal benefits comparable to those achievable with a 10%–20% increase in surgeon capacity.

With population-average information, cardiologists are inclined to refer all patients to elite surgeons. However, armed with patient-specific information, they will appropriately refer some patients to non-elite surgeons, because these surgeons' quality is comparable to that of elite surgeons for some patients. This will not only help spread the workload across surgeons but also help non-elite surgeons improve their skills by giving them some patient volume.

To be effective, a metric must capture patients' concerns about quality, while also being understandable to patients. Experimental research will be needed to determine how best to strike such a balance via effective design of the web sites that distribute patient-specific outcome information. One way to provide patient-specific information is via an interactive web site that first asks a patient to enter his/her demographics, conditions and weights on different quality metrics, and then presents comparative health care information customized to that patient. Existing web sites such as the Online STS Adult Cardiac Surgery Risk Calculator and Heart Risk Calculator allow patients to enter their demographics and conditions but do not compare outcomes of different providers. In contrast, existing hospital rating systems such as Hospital Compare and US News do not require input from patients as they compare providers for only an average patient. New web sites with both features would allow patients to find more tailored health care information. These sites could also be integrated with health care portals to provide context relevant information. For example, the primary care division at Massachusetts General Hospital offers online decision aids targeted at patient needs when patients log into their health portals.[19]

Posting patient-specific information online will only lead to better patient decisions if patients understand it. Experimental research is needed to determine how to strike the right balance between information accuracy and simplicity. Today's rankings based on population-average estimates of individual statistics (e.g., mortality rate) are simple to understand but not at all accurate as gages of provider quality. A customized ranking based on patient-specific estimates of custom weighted QALE metrics is highly accurate but may not be comprehensible to many people.

If research shows that even well-designed web sites do not make patient-specific outcome information understandable to average patients, an alternate channel for disseminating this information is health care providers. For example, a site could be designed to be used by primary care physicians who refer patients to surgeons.

Payers can also play an important role in influencing patient choices. Although hospitals with elite surgeons tend to charge a premium on the surgical procedure itself, their lifetime treatment costs are often lower due to avoidance of complications and readmissions (Wang et al. 2018). This implies that payers have incentive to use of patient-specific information to guide patients (via reduced co-pays for patients and/or

[19] http://www.massgeneral.org/decisionsciences/research/About_Shared_Decision_Making.aspx

value-based compensation of hospitals and surgeons) to providers that offer both better clinical outcomes and lower lifetime costs.

Finally, government agencies and employers can make use of patient-specific outcome data to identify the best providers for groups of patients and encourage patients to choose the best provider for them by subsidizing travel costs. An example of such a travel subsidy is the Healthcare Travel Costs Scheme in the UK, which was set up to provide financial assistance to patients who do not have a medical need for ambulance transport, but who require assistance with their travel costs.[20] Employer examples include Walmart and Lowes, who joined Pacific Business Group On Health to subsidize employees' costs of traveling and lodging when treated at Centers of Excellence for high risk procedures such as heart surgery or knee/hip replacement.[21]

One potential limitation of the approach in this study is that it implicitly assumes patients have the same set of elite surgeons. In some settings, it may be possible that a surgeon performs well on some patients (e.g., young patients) but not so well on other patients (e.g., old patients). In theory, it is possible to address this issue by interacting surgeon dummies with patient characteristics (e.g., age, gender, race, comorbidities). However, when there is a large number of patient characteristics, this approach will create computational burdens as well as generate many statistically insignificant estimates. Future research is needed to characterize such heterogeneous surgeon effects. Another potential limitation is that, since patients are not randomly assigned to providers, our estimates may be biased. There may be characteristics (e.g., echocardiogram) that providers and/or patients themselves observe but we as researchers cannot, which may affect provider/patient selection and patient outcome. To assess whether such biases may affect our conclusion, we use distance-based instruments for surgical volumes and find that our conclusion regarding the benefits from using patient-specific information still holds. Even though this approach does not fully address potential selection issues, it shows that our conclusion is likely robust to potential selection biases.

[20] http://www.nhs.uk/NHSEngland/Healthcosts/Pages/Travelcosts.aspx

[21] Walmart, Lowe's and Pacific Business Group On Health Announce A First Of Its Kind National Employers Centers Of Excellence Network. Walmart News & Views. October, 2013.

CHAPTER 3

An Instrumental Variable Tree Approach for Detecting Heterogeneous Treatment Effects in Observational Studies

3.1 Introduction

The big data revolution presents many opportunities for organizations to personalize their offerings to the heterogeneous needs of their stakeholders. For example, online retailers can use consumer search and clickstream data to understand how consumers respond differently to advertisements; healthcare providers can use clinical and mobile health data to understand how patients respond differently to drugs or treatments; managers can use employee activity and performance data to understand how employees respond differently to reward programs; educators can use data from online learning platforms to understand how students respond differently to pedagogical methods. By understanding heterogeneous responses of different subjects and the factors that drive heterogeneity, organizations can personalize products and services.

A standard approach to analyzing heterogeneous treatment effects is to partition subjects into subgroups based on their features such as demographics and then estimate conditional average treatment effect for each subgroup. For example, a researcher can partition subjects based on gender and estimate average treatment effects separately for the female and male subgroups. However, there are several problems with this approach. First, it is unclear which features should be used for partitioning. Given a large number of observations and features, it is almost always possible to find a feature that appears to be associated with treatment effect heterogeneity. This may lead to dubious results and willful manipulation if a researcher selectively reports results. Second, if there are limited data, it is difficult to partition subjects into subgroups or detect significant differences of treatment effects due to a reduced sample size. Third, treatment effect heterogeneity might still exist within a given subgroup. For example, in the female subgroup, it is possible for young female and old female to respond differently to the same treatment.

A more sophisticated approach interacts features with the treatment in a regression model. One problem with this approach is that it assumes the effects of different features are linearly additive. Another problem is that, if there are a large number of features but a limited number of observations, it can be challenging to obtain reliable estimates of many interaction terms. A third problem with this approach is that it does not explicitly identify subgroups of subjects that have heterogeneous treatment effects. As a result, this approach does not offer a clear reference group with which to compare effects. An alternative approach is to interact predefined subgroups instead of features with the treatment. But how many subgroups to have and which subjects should be in each subgroup are precisely the questions we seek to answer.

Advances in the development of machine learning techniques provide new insights into subgroup analysis. The regression tree approach (Breiman et al. 1984) partly addresses these challenges by recursively partitioning subjects into smaller subgroups such that subjects in the same subgroup have similar outcomes and those in different subgroups have different outcomes. It uses cross-validation to decide on the complexity of a tree. However, the regression tree approach cannot be used for our purpose, because the main purpose of a regression tree is to predict outcomes whereas our purpose is to predict treatment effects.

Recently, Athey and Imbens (2016) proposed a causal tree for analyzing heterogeneous treatment effects when objects are randomly assigned to receive a treatment. This approach partitions subjects into subgroups such that subjects in the same subgroup have similar treatment effects and those in different subgroups have different treatment effects. In randomized controlled experiments, because treatment assignment is not confounded with features, it is straightforward to estimate the treatment effect using the average outcome difference between the treatment and control groups.

While randomized controlled experiments are ideal for causal inference, sometimes, it is unethical, unaffordable or impossible to carry out large-scale randomized controlled experiments. Despite the difficulty in meeting the strict inclusion and exclusion criteria during participant recruitment, randomized controlled experiments are often expensive and take a long time to complete. For example, in context of health care, randomly assigning patients to a treatment that is potentially harmful raises serious ethical concerns. Ethical concerns may also arise when a patient is prevented from receiving a better and more suitable treatment. Finally, there are cases when randomized controlled experiments are infeasible due to legal issues or unavailability of participants. For example, it would be illegal to recruit adolescents to study the impact of smoking at a young age on the development of a lung cancer.

In the absence of randomized controlled experiments, researchers and policy makers have turned to observational data. A major problem with observational data is that there are potential endogeneity issues because observational data often do not include all features that affect treatment assignment and outcome. In the context of health care, there are patient features such as diagnosis results that physicians or patients themselves observe but we as researchers do not. If these features affect both the treatment assignment and medical outcome, simply taking the average outcome difference between the treatment and control groups will lead to biased estimates of the treatment effects.

A number of studies have assumed that assignment of subjects to the treatment or control group is independent of potential outcomes after controlling for observable features of the subjects (i.e., unconfoundedness assumption) and have used propensity score matching to estimate the treatment effects (see e.g., Hahn, Murray and Carvalho 2017, Powers et al. 2017, Wager and Athey 2018, Xie et al. 2012). However, this approach matches subjects in the treatment group to those in the control group based on only observable features. Hence, it is not guaranteed that subjects in the two groups have similar unobservable features. A number of studies have pointed out that propensity score matching does not properly address the endogeneity issue when the unconfoundedness assumption does not hold (see e.g., Breen et al. 2015, King and Nielsen 2016).

The instrumental variable method has been widely used in the operations management literature to correct for potential endogeneity issues (see e.g., Bartel, Chan and Kim 2016, Chan et al. 2016, Freeman et al. 2016, Ho et al. 2000, KC and Terwiesch 2011, KC and Terwiesch 2012, Kim et al. 2014, Lu et al. 2017, McClellan et al. 1994, Xu et al. 2017). A valid instrument induces changes in treatment assignment but has no independent effect on the outcomes, which allows a researcher to uncover the causal effect of the treatment on the outcome. However, use of the instrumental variable method has been limited to regression models. It has not been applied to tree-based approaches.

We address the gap by developing a new instrumental variable tree (hereinafter referred to as "IV tree") that combines the causal tree approach with the classical instrumental variable method to study heterogeneous treatment effects using observational data. This approach addresses the inability of the causal tree to account for endogeneity and does not rely on the unconfoundedness assumption. It allows researchers to perform heterogeneous treatment effect analysis and, at the same time, correct for potential endogeneity biases with observational data.

3.2 Literature Review

A large body of literature on treatment effect analysis has focused on estimation of the average effect of a treatment (see e.g., Lacy et al 2002, Lieberman et al. 2005, Moss et al. 2003). This literature implicitly assumes that treatment effect is homogenous for individual subjects. But, recognizing that average treatment effect presents only the mean effect of a treatment and does not indicate how an individual subject responds to a treatment, a number of scholars have called for heterogeneous treatment effect analysis (Kent et al. 2007, Kravitz et al. 2004, Mant 1999, Vuik et al. 2016).

A parametric approach for heterogeneous treatment effect analysis describes responses as a function of subject features and the treatment, as well as the interaction of some features with the treatment, in a regression model. Because it is challenging to estimate a large number of interaction variables, a number of studies have used Least Absolute Shrinkage and Selection Operations (LASSO) or LASSO-based methods to reduce the dimension of the problem and to identify features that significantly affect treatment effect heterogeneity (see e.g., Imai and Ratkovic 2013, Signorovitch 2007, Taddy et al. 2015, Tian et al. 2014).

Tree-based approaches have gained increasing popularity in the recent years. The conventional Classification And Regression Trees (CART) method focuses on outcome prediction and cannot be applied directly to analyze heterogeneous treatment effects. Building on the ideas of the CART method, studies in the machine learning and statistics literatures have developed new tree-based approaches to focus specifically on heterogeneous treatment effect analysis (see e.g., Athey and Imbens 2016, Chipman et al. 2010, Hothorn et al. 2006, Su et al. 2009, Wager and Athey 2018, Zeileis et al. 2008).

Both the LASSO- and tree-based approaches discussed above are designed for randomized experimental studies and cannot be used for observational studies where treatments may be endogenously determined. A few studies have made the unconfoundedness assumption and used propensity score matching to homogenize subjects in the treatment and control groups based on observable features (Hahn, Murray and Carvalho 2017, Powers et al. 2017, Wager and Athey 2018, Xie et al. 2012). However, the unconfoundedness assumption is not guaranteed to be satisfied in observational studies (Breen et al. 2015, Xie et al. 2012). For example, in the context of health care, there are often features that patients or health care providers observe, but we as researchers do not (Listl et al. 2016, Velentgas et al. 2013). If these features affect both outcome and treatment assignment, matching may not solve the issue.

The endogeneity issue can be corrected when there is an instrumental variable, which correlates with the treatment assignment but has no direct impact on outcomes. The instrumental variable method has been widely used in the operations management literature (see e.g., Bartel, Chan and Kim 2016, Chan et al. 2016, Freeman et al. 2016, Ho et al. 2000, KC and Terwiesch 2011, KC and Terwiesch 2012, Kim et al. 2014, Lu et al. 2017, McClellan et al. 1994, Xu et al. 2017). These studies all recognized potential endogeneity issues in observational studies and used the instrumental variable method to correct for biases. However, they focused only on measuring the average treatment effect. We contribute to this literature by proposing a tree-based approach that incorporates the instrumental variable method to study heterogeneous treatment effects.

In parallel with our work, Athey et al. (2019) developed a generalized random forest approach to partition observations into subgroups based on a set of local moment conditions. They noted that their approach can be used to estimate heterogeneous treatment effects via instrumental variables. However, there are three key differences between their method and ours. First, their approach uses a gradient-based approximation for tree splitting, which results in a loss of efficiency, whereas our approach uses the exact loss function for tree splitting. Second, their splitting rule considers only the mean of treatment effects, which leads to unstable trees, whereas our splitting rule considers both the mean and variance of estimated treatment effects to balance relevance (smaller subgroup size) with reliability (less estimation noise). And third, because their approach partitions observations based on moment conditions, which require all exogenous features to be orthogonal to the error term, the splitting criterion of the generalized random forest is influenced by features directly affecting outcomes, whereas our approach focuses only on estimating treatment effects and thus is more resistant to irrelevant features.

3.3 Problem Formulation and the IV Tree Approach

Suppose we have access to N independently and identically distributed observations, indexed by $i = 1, ..., N$, each of which consists of a d-dimensional feature vector $X_i =$

$\{X_{i1}, X_{i2}, ..., X_{id}\}$, an outcome variable Y_i, and a binary variable $T_i \in [0,1]$ indicating whether subject i received the treatment or not. In addition to X_i, there may be unobservable features that affect both outcome variable and treatment assignment, which creates an endogeneity issue. The problem is to determine if there exist distinct subgroups of subjects across which treatment effects are heterogeneous, and if so, how to estimate the treatment effect for each subgroup.

3.3.1 Regression Approach

When the dimension of the feature vector is small, a parametric regression model may be used for heterogeneous treatment effect analysis. To illustrate this, consider a simple example with a three-dimensional vector of binary features. That is, $X_i = \{X_{i1}, X_{i2}, X_{i3}\}$, where $X_{ij} \in [0,1]$ for $j = \{1,2,3\}$. Because we do not know a prori which features and how their interactions affect the treatment effect, we include all features and their interactions with the treatment in a regression model:

$$\begin{aligned} Y_i \quad = \quad & \alpha_0 + \alpha_1 X_{i1} + \alpha_2 X_{i2} + \alpha_3 X_{i3} + \alpha_4 X_{i1} X_{i2} + \alpha_5 X_{i1} X_{i3} + \alpha_6 X_{i2} X_{i3} + \alpha_7 X_{i1} X_{i2} X_{i3} \\ & + \beta_0 T_i + \beta_1 X_{i1} T_i + \beta_2 X_{i2} T_i + \beta_3 X_{i3} T_i + \beta_4 X_{i1} X_{i2} T_i + \beta_5 X_{i1} X_{i3} T_i + \beta_6 X_{i2} X_{i3} T_i \\ & + \beta_7 X_{i1} X_{i2} X_{i3} T_i + \epsilon_i \end{aligned} \tag{3.1}$$

The parameters of interest are β_0, β_1, ..., and β_7. To understand if the treatment effect is heterogeneous, we need to predefine subgroups and compare the joint distribution of βs related to the predefined subgroups. For example, if we want to compare the treatment effects for Subgroup 1 with feature $\{X_{i1} = 1, X_{i2} = 1, X_{i3} = 1\}$ and Subgroup 2 with feature $\{X_{i1} = 1, X_{i2} = 1, X_{i3} = 0\}$, we need to test if $\beta_0 + \beta_1 + \beta_2 + \beta_3 + \beta_4 + \beta_5 + \beta_6 + \beta_7$ (i.e., sum of Subgroup 1 related coefficients) is significantly different from $\beta_0 + \beta_1 + \beta_2 + \beta_4$ (i.e., sum of Subgroup 2 related coefficients), which is equivalent to test if $\beta_3 + \beta_5 + \beta_6 + \beta_7$ is significantly different from zero.

This approach presents four major challenges when the dimension of the feature vector is large. First, the number of terms in the regression model increases exponentially with the dimension of the feature vector. In the case of N binary features, there are a total of 2^N unique combinations of features that could affect the treatment effect. There are even more terms when some of the features are continuous instead of binary. Second, this approach does not explicitly identify subgroups of subjects that have heterogeneous treatment effects. That is, it is unclear how many distinct subgroups there are and which subjects belong to the same subgroup. Third, estimates of the βs will be biased if there are unobservable features that affect both the outcome variable and treatment assignment. It is impractical to address the endogeneity issue using the IV approach in this case, because this approach requires a large number of instruments (e.g., 2^N in the case of binary features). Fourth, this parametric approach assumes that the effects of different features are linearly additive, whereas in reality the features may interact in a complicated and nonlinear way. To address these challenges, we use a tree-based approach as what we will discuss next.

3.3.2 Tree-Based Approaches

In the machine learning literature, a tree-based approach is a non-parametric method that recursively partitions subjects (based on one feature at a time) into subgroups such that those in the same subgroup have similar parameters of interest (e.g., outcome or treatment effect) and those across different subgroups have different parameters of interest. It is called a tree-based approach because the set of splitting rules used to partition subjects can be summarized in a tree. The most well known tree-based approach is the regression tree, which partitions subjects into heterogeneous subgroups based on outcomes. The regression tree is not suitable for heterogenous treatment effect analysis, because the features that affect the outcome variable may be different from those affect the treatment effect.

The causal tree approach extends the regression tree for heterogeneous treatment effect analysis when subjects are randomly assigned to receive the treatment or control. Because there is no confounding features that affect both the outcome variable and treatment assignment, the treatment effect β_{l_j} for subgroup l_j with feature x_{l_j} (i.e., $x_{l_j} = \{x_k : X_{ik} = x_k, \forall i \in l_j\}$) can be estimated using the average outcome difference between the treatment and control groups (denoted as $\overline{y}_{1l_j}$ and $\overline{y}_{0l_j}$, respectively). That is

$$\hat{\beta}_{CT}(x_{l_j}) \;\;=\;\; \overline{y}_{1l_j} - \overline{y}_{0l_j} = \tfrac{1}{N_{1l_j}} \textstyle\sum_{i \in l_j, T_i = 1} Y_i - \tfrac{1}{N_{0l_j}} \textstyle\sum_{i \in l_j, T_i = 0} Y_i, \qquad (3.2)$$

where N_{1l_j} and N_{0l_j} denote the numbers of subjects (in subgroup l_j) that received the treatment and control, respectively, and T_i indicates whether subject i received the treatment or not.

Because observations in subgroup l_j are independently and identically distributed, the variance of $\hat{\beta}_{CT}(x_{l_j})$ can be estimated as

$$Var[\hat{\beta}_{CT}(x_{l_j})] \;=\; \frac{S^2_{1l_j}}{N_{1l_j}} + \frac{S^2_{0l_j}}{N_{0l_j}}, \tag{3.3}$$

where $S^2_{1l_j}$ and $S^2_{0l_j}$ are within-group variances of outcomes of the subjects that received the treatment and control, respectively.

The causal tree uses the "honest" approach (see e.g., Green and Kern 2010, Heller et al. 2009) to randomly divide the data into two parts – one part (denoted as S^{tr}) for training the tree model and the other part (denoted as S^{es}) for estimating the treatment effects. Let N^{tr} and N^{es} denote the sizes of training and estimation samples, respectively. The causal tree starts at the root of the tree where all subjects are in the same group and recursively partitions subjects into two smaller subgroups based on the feature that reduces $-(1/N^{tr})\sum_j \hat{\beta}_{CT}(x_{l_j})^2 + (1/N^{tr} + 1/N^{es})\sum_j Var[\hat{\beta}_{CT}(x_{l_j})]$ by the greatest amount. The process is repeated until a stopping criterion is reached. It then prunes the initial large tree to obtain a set of subtrees and uses cross-validation to select the best subtree.

When the treatment is not randomly assigned, as is often the case in observational studies, taking the difference of the average outcomes of the treatment and control groups, $\frac{1}{N_{1l_j}}\sum_{i\in l_j, T_i=1} Y_i - \frac{1}{N_{0l_j}}\sum_{i\in l_j, T_i=0} Y_i$, will lead to biased estimates of the treatment effects. As a result, the causal tree may partition subjects into incorrect subgroups and provide biased estimates of the treatment effects. We propose the IV tree approach to address this issue.

3.3.3 The IV Tree Approach

To describe the IV tree approach, we let ε_i denote unobservable features that correlate with both the outcome variable (i.e., $Cov(Y_i, \varepsilon_i) \neq 0$) and treatment assignment (i.e., $Cov(T_i, \varepsilon_i) \neq 0$), and ξ_i denote an idiosyncratic error. The potential outcome Y_i of subject i in candidate subgroup l_j can be written as

$$Y_i \;=\; \alpha_i(X_i) + \beta_i(X_i)T_i + \epsilon_i, \forall i \in l_j, \tag{3.4}$$

where $\alpha_i(X_i)$ and $\beta_i(X_i)$ are functions that describe how features affect mean outcomes and treatment effects, respectively, and $\epsilon_i = \varepsilon_i + \xi_i$. For notational convenience, we will suppress the dependence of α_i and β_i on X_i in cases where it is unambiguous.

In general, we cannot estimate fully personalized treatment effects for individual subjects, because all statistical parameters of interest (e.g., β_i in our study) can be computed only at the subgroup level. A common approach to addressing this issue is to temporarily ignore individual-level heterogeneity within a subgroup and focus on conditional average treatment effects by assuming that observations in the same subgroup have the same treatment effect (Athey and Imbens 2016, Wager and Athey 2018, Xie et al. 2012). As what we will show later, the instrumental variable tree approach proposed in this study recovers some degree of individual-level heterogeneity by partitioning observations into subgroups with more homogeneous features as sample size increases.

Suppose there exists a variable Z_i that correlates with the treatment assignment, $Cov(Z_i, T_i) \neq 0$, (i.e., satisfying the relevant condition) but does not correlate with the error term, $Cov(Z_i, \epsilon_i) = 0$, (i.e., satisfying the exogeneity condition). We can then use the variable Z_i as an instrument for the treatment dummy T_i (Greene 2003). Given a subgroup l_j with feature x_{l_j} and average treatment effect β_{l_j}, the treatment effect β_{l_j} can be estimated as

$$\hat{\beta}_{IV}(x_{l_j}) \quad = \quad \tfrac{Cov(Y_i, Z_i | i \in l_j)}{Cov(T_i, Z_i | i \in l_j)}. \tag{3.5}$$

The variance of $\hat{\beta}_{IV}(x_{l_j})$ is

$$Var[\hat{\beta}_{IV}(x_{l_j})] \quad = \quad \tfrac{Var(\epsilon_i | i \in l_j)}{N_{l_j} Var(T_i | i \in l_j)[Cor(T_i, Z_i | i \in l_j)]^2}, \tag{3.6}$$

where N_{l_j} is the number of subjects in subgroup l_j. A consistent estimator of $Var(\epsilon_i | i \in l_j)$ is $(\sum_{i \in l_j} \hat{\epsilon}_i^2)/(N_{l_j} - 2)$, where $\hat{\epsilon}_i = Y_i - \hat{\alpha}_i - \hat{\beta}_i T_i$, and $\hat{\alpha}_i$ and $\hat{\beta}_i$ are the IV estimates. Note that $N_{l_j} - 2$ is used for the degrees of freedom correction.

The challenge now is to partition subjects into proper subgroups such that subjects in each subgroup have similar treatment effects. To construct a tree-based algorithm, we begin with a root in which all subjects are in the same node. The tree grows by selecting a node (called a "parent node") and partitioning subjects into two nodes (called "child nodes") based on differences in a feature. Partitions could be made on a binary feature such as gender or a non-binary feature such as age by using a cutoff value.

The key to the performance of a tree-based approach is the splitting rule, i.e., how to choose a feature to split on and, if the feature is not binary, how to choose a cutoff value. As the final goal of a tree is to predict treatment effects for new subjects, we grow the IV tree in a way that minimizes the mean-squared error of estimated treatment effects. Given a testing sample S^{te} used to evaluate tree performance and an estimation sample S^{es} used to estimate treatment effects, we let $\beta^{te}(X_i)$ and $\hat{\beta}^{es}_{IV}(X_i)$ denote the true and estimated treatment effects for subject i with feature X_i, and define the mean-squared error of a tree π as

$$MSE(S^{te}, S^{es}) \;\; = \;\; \tfrac{1}{N^{te}} \textstyle\sum_{i \in S^{te}} [\beta^{te}(X_i) - \hat{\beta}^{es}_{IV}(X_i)]^2. \qquad (3.7)$$

The loss function, $EMSE(S^{te}, S^{es})$, is equal to the expectation of $MSE(S^{te}, S^{es})$ over testing and estimation samples in the honest approach. Let $\beta(X_i)$ denote the conditional average treatment effects for subject $i \in l_j$. By the law of conditional expectation and observing that $E_{S^{te}}[\beta^{te}(X_i)^2]$ does not depend on the tree structure, we have (see Appendix B.1 for more details)

$$EMSE(S^{te}, S^{es}) \;\; = \;\; -E_{X_i}[\beta(X_i)^2] + E_{X_i, S^{es}}[Var(\hat{\beta}^{es}_{IV}(X_i))]. \qquad (3.8)$$

where the first and second terms relate to the mean and variance of estimated treatment effects, respectively.

The first term, $E_{X_i}[\beta(X_i)^2]$, can be estimated using the square of the estimated treatment effects from the training sample, $\hat{\beta}^{tr}_{IV}(X_i)^2$, minus an estimate of its variance weighted by the fractions of observations (of the training sample) in the terminal nodes: $\hat{E}_{X_i}[\beta(X_i)^2] = \sum_j (N_{l_j^{tr}}/N^{tr}) \times [\hat{\beta}^{tr}_{IV}(x_{l_j})^2 - Var(\hat{\beta}^{tr}_{IV}(x_{l_j}))]$, where $\hat{\beta}^{tr}_{IV}(x_{l_j})$ and $Var(\hat{\beta}^{tr}_{IV}(x_{l_j}))$ are estimated using the IV method described earlier.

The second item, $E_{X_i, S_{es}}[Var(\hat{\beta}^{es}_{IV}(X_i))]$, can be calculated as a weighted sum of within-group variances, where the weights are the fractions of observations (of the estimation sample) in the terminal nodes. Within-group variances can be calculated using estimates of $Var(\epsilon_i | i \in l_j)$, $Var(T_i | i \in l_j)$, and $Cov(T_i, Z_i | i \in l_j)^2$ from the training sample.

3.3.4 IV Tree Algorithm

In this section, we describe an algorithm for constructing an IV tree. The algorithm consists of three major steps: (1) growing an initial large tree using the splitting rule discussed earlier; (2) recursively pruning the initial large tree based on the weakest links to obtain a set of subtrees; and (3) selecting the best subtree via cross-validation and estimating the treatment effect for each subgroup using the honest approach.

3.3.4.1 Growing An Initial Large Tree

To grow an IV tree, we evaluate all nodes, candidate features and all possible cutoff values for each feature and partition subjects in a selected parent node into two child nodes based on the feature and cutoff point that reduces $EMSE(S^{te}, S^{es})$ by the greatest amount. We then treat each child node as a parent node and repeat the process of partitioning until reaching a stopping criterion (e.g., each node must include at least 30 observations). This recursive partitioning process leads to an initial large tree.

3.3.4.2 Prune the Tree Based on the Weakest Links

A very large tree (e.g., each terminal node having a single observation) will have a poor out-of-sample goodness of fit due to overfitting of the training data. A very small tree (e.g., a single node including all observations) may also have a poor out-of-sample goodness of fit, because it does not capture the important underlying structure of the data. Our objective is to find the right-sized subtree, defined as a tree that can be obtained by pruning the initial large tree (including no pruning), that provides the best out-of-sample goodness of fit. We achieve this objective by first pruning the initial large tree to obtain a set of subtrees and then using cross-validation to select the best subtree.

We identify the set of subtrees using the weakest link pruning approach proposed by Breiman et al. (1984).[1] To describe this pruning approach, we let M_{π_i} denote the complexity (i.e., number of terminal nodes) of tree π_i, h be an internal node of the initial large tree π_0, and $\pi_1(h)$ be a subtree of π_0 after deleting all branches connecting to h. The weakness of node h is calculated as $W_h(\pi_0) = (EMSE_{\pi_0} - EMSE_{\pi_1(h)})/(M_{\pi_0} - M_{\pi_1(h)})$. By comparing the weakness of all internal nodes, we identify the weakest link of π_0 as the node with the largest value of $W_h(\pi_0)$, i.e., $h^* = argmax_h\{(EMSE_{\pi_0} - EMSE_{\pi_1(h)})/(M_{\pi_0} - M_{\pi_1(h)})\}$ and use it for pruning. We then prune tree $\pi_1(h^*)$ in the same way as we did for π_0. Repeating this pruning process leads to a series of subtrees, $\pi_0 \subset \pi_i \subset \pi_2 \subset \cdots \subset \pi_N$, where π_N is the single-node tree with all subjects in the node.

3.3.4.3 Select the Best Subtree and Estimate Treatment Effects

We choose the best subtree using five-fold cross-validation.[2] That is, to evaluate the performance of subtree π_i, we randomly divide the training data into five equal folds. Each time we hold out one fold of the data for validation and use the other four folds for training. Let $EMSE_{\pi_i}(S^j, S^{es})$ denote the expected mean-squared error of π_i when

[1] This approach is also called cost complexity pruning, as the weakest link is the node associated with the largest ratio of the change in cost to the change in complexity.

[2] We use five-fold instead of ten-fold cross-validation, because we follow the honest approach that uses one half of the data for training and the other half for estimation.

the jth fold of the data is held out. The average expected mean-squared error of π_i from cross-validation is $AEMSE(\pi_i) = \frac{1}{5}\sum_{j=1}^{5} EMSE_{\pi_i}(S^j, S^{es})$.

The best subtree is defined as the one that minimizes the average expected mean-squared error $\pi^* = argmin_{\pi_i} AEMSE(\pi_i)$. After choosing the best subtree, it is straightforward to estimate treatment effects for terminal nodes using the estimation sample with the IV method. Subjects in the same terminal node are expected to have the same treatment effect.

The algorithm for constructing an IV tree is summarized below:

IV Tree Algorithm

1. Start with the root node where all subjects are in the same group $L = \{S^{tr}\}$. For each subgroup $l_0 \subset L$ that is not a terminal node, do the following:

   a. Calculate the loss function $EMSE_{l_0}(S^{te}, S^{es})$;

   b. For each feature X (and a possible cutoff value if X is continuous), do the following:

   - partition subjects in l_0 into two subgroups, l_1 and l_2, based on X (and the cutoff value),
   - calculate the loss functions $EMSE_{l_1}(S^{te}, S^{es})$ and $EMSE_{l_2}(S^{te}, S^{es})$,
   - if $EMSE_{l_0}(S^{te}, S^{es}) \times N_{l_0} \geq EMSE_{l_1}(S^{te}, S^{es}) \times N_{l_1} + EMSE_{l_2}(S^{te}, S^{es}) \times N_{l_2}$ and $N_{l_1}, N_{l_2} \geq 30$, replace l_0 with l_1 and l_2; otherwise, let l_0 be a terminal node;

2. Start with the initial large tree Π obtained from Step 1. For each subtree $\pi_0 \subset \Pi$ that is not a single node, do the following:

   a. Calculate $EMSE_{\pi_0}(S^{te}, S^{es})$ and tree complexity M_{π_0} (i.e., number of terminal nodes);

   b. For each internal node $h \in \pi_0$, delete all branches connecting to h to obtain subtree $\pi_1(h)$, and calculate $EMSE_{\pi_1(h)}(S^{te}, S^{es})$ and tree complexity $M_{\pi_1(h)}$;

   c. Identify the weakest link $h^* = argmax_h\{(EMSE_{\pi_0} - EMSE_{\pi_1(h)})/(M_{\pi_0} - M_{\pi_1(h)})\}$, prune π_0 based on h^*, and replace π_0 with $\pi_1(h^*)$;

3. Randomly divide the training sample into five folds:

a. For each fold j, do the following:

- hold out fold j and use the remaining four folds for tree growing and pruning,
- use fold j for cross-validation and calculate $EMSE_{\pi_i}(S^j, S^{es})$;

b. Calculate $AEMSE(\pi_i) = \frac{1}{5}\sum_{j=1}^{5} EMSE_{\pi_i}(S^j, S^{es})$, and select the best subtree $\pi^* = argmin_{\pi_i} AEMSE(\pi_i)$;

4. Estimate the treatment effect for terminal node l_j using $\hat{\beta}_{IV}^{es}(x_{l_j})$.

3.3.5 Asymptotic Properties

Our analysis of the asymptotic properties of the IV tree builds on the ideas developed by Breiman et al. (1984), who studied the asymptotic properties of the regression tree method. In this section, we state a few assumptions that guarantee the consistency of the proposed estimator. All proofs of the lemma, theorem, and corollary are provided in Appendix B.2.

Assumption 1: Let x_{l_j} denote the feature of subgroup l_j (i.e., $x_{l_j} = \{x_k : X_{ik} = x_k, \forall i \in l_j\}$) and $d(x_{l_j})$ denote the largest dissimilarity between any two subjects in subgroup l_j. That is, $d(x_{l_j}) = \sup\limits_{i,j \in l_j} |X_i - X_j|$, where $|X| = (x_1^2 + x_2^2 + ... + x_d^2)^{1/2}$. We assume that $\lim\limits_{N\to\infty} d_N(x_{l_j}) = 0$ in probability.

This assumption has been made in a number of studies (see e.g., Breiman et al. 1984, Chapter 12.2) for proving the asymptotic property of tree-based approaches. It states that the dissimilarity between any two subjects in a terminal node approaches zero as the sample size increases to infinity. Intuitively, a larger sample size allows a tree to be more capable of detecting treatment effect heterogeneity across subjects and split more on feature space, leading to subgroups of subjects with more homogeneous features.

Example: Consider observations with two dimensions of binary features, {male, female} and {young, old}, and a tree that partitions subjects into two subgroups based on gender. Then the features of the two subgroups are {male} and {female}, respectively. The dissimilarity of a subgroup equals one if it includes both young and old patients. As the overall sample size increases, we now consider a larger tree that partitions subjects into four subgroups based on both gender and age. Then the features of the four subgroups are {male, young}, {female, young}, {male, old}, and {female, old}, respectively, and the dissimilarity of each subgroup equals zero.

Assumption 2: The treatment effect β_i is a continuous function of observable feature X_i. That is, $\lim\limits_{X_i \to X_j} \beta_i(X_i) = \beta_j(X_j)$.

First, we assume that the treatment effect is a function of observable features. Because it is impossible to partition subjects based on unobservable features, all studies performing subgroup analysis are implicitly making this assumption. Second, we assume that the function linking treatment effect and observable features is continuous. This implies that, as the dissimilarity between two subjects approaches zero, the treatment effect difference between the two subjects approaches zero.

Example: Suppose the effect of warfarin on reducing blood clots is a function of patient weight and height only. We assume that, as a patient's weight or height changes, the treatment effect of warfarin remains the same or changes continuously. This assumption is reasonable as we do not expect a jump in the treatment effect of warfarin when a patient has a slight increase in height or decrease in weight.

With Assumptions 1 and 2, we can state the following lemma:

Lemma 1: Suppose Assumptions 1 and 2 hold. For all $i \in l_j$, there exists β_{l_j} such that $\beta_i \xrightarrow{p} \beta_{l_j}$, where $\beta_{l_j} = \beta_i(\lim\limits_{N_{l_j} \to \infty} x_i) = \beta_i(x_{l_j})$. Also, $Cov(\beta_i T_i, Y_i) \xrightarrow{p} Cov(\beta_{l_j} T_i, Y_i)$.

Assumption 3: Let V_i denote the product of the IV and error term (i.e., $V_i = Z_i \epsilon_i$) and Q_{l_j} denote the expected product of the IV and treatment dummy (i.e., $Q_{l_j} = E[Z_i T_i], \forall i \in l_j$). We assume that both V_i and Q_{l_j} are bounded.

The first part of the assumption ensures that the moment-generating function of V_i is bounded. The second part of the assumption is a standard assumption in the literature to prove the consistency of IV estimators (see e.g., Greene 2003, Chapter 5.4). In practice, because the IV and outcome measure in observational studies are usually bounded, it is reasonable to assume that V_i and Q_{l_j} are bounded.

Example 1: Lee et al. (2017) studied the impact of operative time (i.e., the empirical treatment) on graft survival (i.e., the outcome variable) after liver transplantation. They used average risk-adjusted operative time of the most recent cases performed by the same surgeon as an instrument for operative time of the focal patient. Our assumption is valid in this case because both operative time and post-transplant survival are bounded.

Example 2: Gowrisankaran and Town (1999) compared the morality rate (i.e., the outcome variable) of different hospitals (i.e., the empirical treatments) for pneumonia care in Southern California. They used the travel distance from a patient to a hospital as an instrument for the hospital dummy, as travel distance correlates with the choice of a hospital but does not correlate with patient sickness. Our assumption is also valid in this case because both distance and mortality rate are bounded.

Assumption 4: Let N and N_{l_j} denote the overall sample size and number of subjects in subgroup l_j, respectively. There exists a sequence of positive numbers k_N such that $\lim_N k_N = \infty$ and $N_{l_j}/logN \geq k_N$.

This assumption states that the number of subjects in a terminal node increases as the overall sample size increases and the rate of increase (i.e., dN_{l_j}/dN) is larger than k_N/N. We make this assumption to ensure that there are sufficiently many observations in each terminal node as the total number of observations increases. This is to guarantee that the estimator of the treatment effect for any subgroup is asymptotically consistent.

Example: Consider observations with two dimensions of features, {male, female} and {young, old}, and a tree that partitions subjects into four subgroups, {male, young}, {female, young}, {male, old}, and {female, old}. This assumption states that, as the overall sample size increases (e.g., $N = \{10, 100, 1000, ...\}$), there exists a sequence of positive numbers (e.g., $k_N = \{1, 2, 3, ...\}$) such that the number of subjects in each subgroup increases accordingly (e.g., $N_{l_j} \geq logN \times k_N = \{1, 4, 9, ...\}, \forall j = 1, 2, 3, 4$).

With Assumptions 3 and 4, we can state the following lemma:

Lemma 2: Let a and b denote the lower and upper bounds of V_i (i.e., $a \leq V_i \leq b$) and $\psi_i(t)$ denote the moment-generating function of V_i (i.e., $\psi_i(t) = E[e^{tV_i}]$). If Assumptions 3 and 4 hold, we have $\psi_i(t) \leq e^{\frac{t^2(b-a)^2}{8}}$. Also, for any $w > 0$, we have $Pr(\overline{V} \geq w) \leq N^{-7w^2 k_N/[8(b-a)^2]}$, where k_N is a sequence of positive numbers with $\lim_N k_N = \infty$.

To prove consistency of the proposed estimator, we write the potential outcome of subject i in subgroup l_j as $Y_i = \alpha_i + \beta_i T_i + \epsilon_i$, where β_i is the treatment effect for subject i. Our goal is to prove that $\hat{\beta}_{IV}(x_{l_j}) = Cov(Y_i, Z_i|i \in l_j)/Cov(T_i, Z_i|i \in l_j)$ is a consistent estimator for β_i, for all $l_j \in \pi$ and $i \in l_j$. That is, we want to show that $\hat{\beta}_{IV}(x_{l_j}) \xrightarrow{p} \beta_i, \forall l_j \in \pi, \forall i \in l_j$. Because $\beta_i \xrightarrow{p} \beta_{l_j}$ (from Lemma 1), it suffices to show that $\hat{\beta}_{IV}(x_{l_j}) \xrightarrow{p} \beta_{l_j}, \forall l_j \in \pi, \forall i \in l_j$. We now state the main theorem of this paper:

Theorem 1: Under Assumptions 1–4, $\hat{\beta}_{IV}(x_{l_j})$ is a consistent estimator of β_i for all $l_j \in \pi$ and $i \in l_j$. That is

$$\max_{l_j \in \pi} \sup_{i \in l_j} |\hat{\beta}_{IV}(x_{l_j}) - \beta_i| \xrightarrow{p} 0$$

in probability as $N \rightarrow \infty$.

When there are endogeneity issues, the causal tree does not provide consistent estimates of treatment effects. We formally state this point in the following corollary:

Corollary 1: Let ε_i and $\hat{\beta}_{CT}(x_{l_j})$ denote unobservable features and estimators of the causal tree, respectively. If $Cov(\varepsilon_i, T_i) \neq 0$, there exists $i \in l_j$ such that $\hat{\beta}_{CT}(x_{l_j}) \overset{p}{\nrightarrow} \beta_i$.

3.4 Performance on Synthetic Data

We conduct simulation studies to assess the performance of the IV tree. The objective of these simulation studies is to understand whether the IV tree effectively corrects for endogeneity biases and, if so, how its performance changes with (1) number of features, (2) feature type (i.e., continuous or binary), (3) error distribution (e.g., normal, exponential, or uniform), (4) model specification (e.g., linear or nonlinear), (5) severity of endogeneity, (6) strength of instrument, and (7) sample size. By constructing the data, we know what the true treatment effects are, so we can compare IV tree estimates with the true treatment effects to see how much they are different from each other.

3.4.1 IV Tree Performance and Comparison with Causal Tree

Because the IV tree extends the causal tree for observational studies, we use the performance of the causal tree as a benchmark for assessing the performance of the IV tree. In the next subsection, we construct an IV forest based on modified IV trees and compare its performance with the generalized random forest.

3.4.1.1 Synthetic Data Construction

To construct the data, we let X_i^α denote the set of features that affect mean outcomes, and X_i^β denote the set of features that affect the treatment effect of subject i. We note that X_i^α and X_i^β may have common elements if there are features that affect both outcomes and treatment effects. Let ε_i and ξ_i denote unobservable features and an idiosyncratic error, respectively. By construction, ε_i correlates with treatment assignment (i.e., $Cor(\varepsilon_i, T_i) \neq 0$) but ξ_i does not (i.e., $Cor(\xi_i, T_i) = 0$). Without loss of generality, we generate outcome Y_i as a function of $\alpha_i(X_i^\alpha)$, $\beta_i(X_i^\beta)$, T_i, ε_i, and ξ_i as follows

$$Y_i = \alpha_i(X_i^\alpha) + \beta_i(X_i^\beta) T_i + \varepsilon_i + \xi_i. \tag{3.9}$$

We consider eight designs (see Table 3.1) to assess the performance of the IV tree under various conditions discussed at the beginning of this section. In Designs 1–2, we consider instances with a total number of features equal to 5, 10, and 20 to understand the performance of the IV tree as the size of the problem increases. Comparing Designs 1 and 2 allows us to understand the performance of the IV tree when the features are binary instead of continuous. Designs 3–6 allow us to understand the performance of the IV tree when the error term has a different distribution or when the model has a different specification. Designs 7–8 allow us to understand how the performance of the IV tree changes with the severity of endogeneity and the strength of the instrument. Finally, we increase the sample size incrementally from 1,000 to 5,000 to understand the small and large sample properties of the IV tree.

Table 3.1: Designs of the Simulation Study

Design	Model	#Features	Notes
1	$Y_i = \sum_{k=1}^{3m} x_{ik} + \sum_{k=1}^{m} x_{ik}T_i + \varepsilon_i + \xi_i$	{5, 10, 20}	$x_{ik}, \varepsilon_i \sim Norm(0,1)$
2	$Y_i = \sum_{k=1}^{3m} x_{ik} + \sum_{k=1}^{m} x_{ik}T_i + \varepsilon_i + \xi_i$	{5, 10, 20}	$x_{ik}, \varepsilon_i \sim Bern(0.5)$
3	$Y_i = \sum_{k=1}^{6} x_{ik} + \sum_{k=1}^{2} x_{ik}T_i + \varepsilon_i + \xi_i$	10	$\xi_i \sim Expo(10)$
4	$Y_i = \sum_{k=1}^{6} x_{ik} + \sum_{k=1}^{2} x_{ik}T_i + \varepsilon_i + \xi_i$	10	$\xi_i \sim Unif(0,1)$
5	$Y_i = \sum_{k=1}^{6} x_{ik} + \Pi_{k=1}^{2} x_{ik}T_i + \varepsilon_i + \xi_i$	10	nonlinear model
6	$Y_i = \sum_{k=1}^{6} x_{ik} + \sum_{k=1}^{2} \mathbb{1}\{x_{ik} > 0\}x_{ik}T_i + \varepsilon_i + \xi_i$	10	nonlinear model
7	$Y_i = \sum_{k=1}^{6} x_{ik} + \sum_{k=1}^{2} x_{ik}T_i + \varepsilon_i + \xi_i$	10	$Corr(\varepsilon_i, T_{1i}) = 0.3$
8	$Y_i = \sum_{k=1}^{6} x_{ik} + \sum_{k=1}^{2} x_{ik}T_i + \varepsilon_i + \xi_i$	10	$Corr(Z_i, T_{1i}) = 0.4$

Note: In Designs 1–2, $m = \{1, 2, 4\}$ indicates the number of features that affect the treatment effect. Unless specified otherwise in the note column, we let $x_i \sim Norm(0,1)$, $\varepsilon_i \sim Norm(0,1)$, $\xi_i \sim Norm(0, 0.01)$, $T_i \sim Bern(0.5)$, $Corr(\varepsilon_i, T_i) = 0.5$, and $Corr(Z_i, T_i) = 0.6$.

In all designs, the features under consideration include both relevant (i.e., those affect mean outcomes, treatment effects, or both) and irrelevant (i.e., those do not affect mean outcomes or treatment effects) features. Some of these features are unobservable to the researchers (e.g., ε_i in Design 1), so they are unavailable for analysis. Features that do not appear in the models (e.g., x_{i7}, x_{i8}, and x_{i9} in Designs 3–8) are irrelevant features. Because in practice we do not know a priori which features are relevant, we include all features for analysis and assess whether the IV tree splits on only features that affect the treatment effect.

3.4.1.2 Performance Metrics

Following existing machine learning literature, we evaluate the performance of the IV and causal trees using coverage rate and mean-squared error of the estimated treatment effects. The coverage rate is defined as the proportion of instances the estimated confidence intervals (e.g., 95%) cover the true values of treatment effects. Mathematically, let β_i denote the true treatment effect for subject i, CI_i denote a confidence interval of the estimated treatment effect for subject i, and $\mathbb{1}\{\beta_i \in CI_i\}$ denote a binary indicator function indicating whether the confidence interval covers the true treatment effect. Then coverage rate is calculated as $Coverage = \frac{1}{N^{te}} \sum_{i=1}^{N^{te}} \mathbb{1}\{\beta_i \in CI_i\}$, where N^{te} denotes the number observations in the test sample.

Coverage rate alone is not sufficient to measure the performance of a tree approach, because an estimator with a large variance may have a high coverage rate. We therefore use mean-squared error as a second performance metric. Let $\hat{\beta}(X_i)$ denote the estimated treatment effect for subject i using a tree approach. The mean-squared error is calculated as $MSE = \frac{1}{N^{te}} \sum_{i=1}^{N^{te}} (\hat{\beta}(X_i) - \beta_i)^2$. The mean squared error measures how much estimated treatment effects deviate from the true treatment effects. For an unbiased estimator, the mean squared error measures the variance of the estimator. For a biased estimator, the mean squared error is affected by both the variance and bias of the estimator.

3.4.1.3 Estimation Results

Table 3.2 summarizes the coverage rate and mean-squared errors of the IV and causal trees. All results are based on a testing sample of size 5,000 and 10 runs of the simulations. We use bootstrap to estimate the 95% confidence interval of the IV tree. Mean-squared error is calculated using the difference between the estimated and true values of treatment effects. We now discuss the performance of the IV tree under various conditions.

Number of Features: As expected, the coverage rate of the IV tree is higher and mean-squared error is smaller when the underlying model is simpler, because there are fewer features affecting the treatment effect. In all scenarios, the IV tree have a better coverage rate and smaller mean-squared error than the causal tree.

Binary Features: Comparing Designs 1 and 2, we see that the IV tree performs better when the underlying models have binary instead of continuous features, due to the natural subgroups implied by binary features. The performance gap between the IV and causal trees is even more substantial when the underlying model has binary

Table 3.2: Comparison of IV Tree (IVT) and Causal Tree (CT)

		Sample Size											
#Features	Approach	1,000	3,000	5,000	1,000	3,000	5,000	1,000	3,000	5,000	1,000	3,000	5,000
		Design 1						Design 2					
		Coverage			MSE			Coverage			MSE		
5	IVT	0.90	0.95	0.97	0.69	0.40	0.31	0.95	0.97	0.92	0.12	0.06	0.05
	CT	0.33	0.13	0.09	1.50	1.23	1.20	0.00	0.00	0.00	1.10	1.04	1.04
10	IVT	0.91	0.91	0.95	1.59	1.15	0.99	0.99	0.99	1.00	0.29	0.11	0.09
	CT	0.74	0.75	0.74	2.19	1.88	1.77	0.23	0.06	0.01	1.32	1.12	1.08
20	IVT	0.83	0.91	0.93	4.62	3.24	2.80	0.89	0.93	0.97	1.00	0.52	0.35
	CT	0.70	0.82	0.85	4.98	4.02	3.77	0.56	0.53	0.51	1.76	1.58	1.47
		Design 3						Design 4					
		Coverage			MSE			Coverage			MSE		
10	IVT	0.95	0.95	0.95	1.27	0.94	0.84	0.95	0.94	0.95	1.25	0.94	0.84
	CT	0.86	0.84	0.65	2.04	1.68	1.62	0.85	0.81	0.71	2.06	1.71	1.58
		Design 5						Design 6					
		Coverage			MSE			Coverage			MSE		
10	IVT	0.87	0.87	0.86	1.36	1.13	1.05	0.98	0.98	0.98	0.74	0.51	0.46
	CT	0.44	0.42	0.39	2.12	1.89	1.77	0.42	0.34	0.27	1.59	1.45	1.38
		Design 7						Design 8					
		Coverage			MSE			Coverage			MSE		
10	IVT	0.95	0.96	0.96	1.26	0.94	0.83	0.93	0.94	0.95	1.85	1.14	0.98
	CT	0.92	0.96	0.97	1.44	1.07	0.96	0.84	0.83	0.68	2.06	1.68	1.59

Note: Designs 1 and 2 have the form $Y_i = \sum_{k=1}^{3m} x_{ik} + \sum_{k=1}^{m} x_{ik}T_i + \varepsilon_i + \xi_i$, where $m = \{1, 2, 4\}$, $x_{ik}, \varepsilon_i \sim Norm(0, 1)$ in Design 1 and $x_{ik}, \varepsilon_i \sim Bern(0.5)$ in Design 2, $\xi_i \sim Norm(0, 1)$, $Corr(\varepsilon_i, T_i) = 0.5$, and $Corr(Z_i, T_i) = 0.6$. Designs 3 and 4 have the same form as Design 1 with $m = 2$ except that $\xi_i \sim Expo(10)$ in Design 3 and $\xi_i \sim Unif(0, 1)$ in Design 4. Designs 5 and 6 have the forms $Y_i = \sum_{k=1}^{6} x_{ik} + \Pi_{k=1}^{2} x_{ik}T_i + \varepsilon_i + \xi_i$ and $Y_i = \sum_{k=1}^{6} x_{ik} + \sum_{k=1}^{2} \mathbb{1}\{x_{ik} > 0\}x_{ik}T_i + \varepsilon_i + \xi_i$, respectively. Designs 7 and 8 have the same form as Design 1 with $m = 2$ except that $Corr(\varepsilon_i, T_i) = 0.3$ in Design 7 and $Corr(Z_i, T_i) = 0.4$ in Design 8. Coverage rate is calculated as the proportion of instances the 95% bootstrap confidence intervals cover the true values of treatment effects. Mean-squared error is calculated using the difference between the estimated and true values of treatment effects. Results are aggregated over 10 runs of the simulations.

features.

Error Distribution: Comparing Designs 3 and 4 with Design 1 with 10 features, we see that the distribution of the error term in the underlying model does not significantly affect the coverage rate or mean-squared error of the IV tree.

Model Specification: As expected, the underlying model specification affects the performance of the IV tree, because the IV tree uses a simple regression model to estimate the treatment effect for each subgroup. Comparing Designs 5 and 6 with Design 1 with 10 features, we see that the performance of the IV tree decreases slightly when features interact with the treatment in a more complex way.

Severity of Endogeneity: Comparing Designs 7 with Design 1 with 10 features, we see that performance of the IV tree remains almost the same and that of the causal tree improves when there is less endogeneity, so the performance gap of the two approaches decreases. However, the IV tree still has a better performance than the casual tree when $Corr(\varepsilon_i, T_i) = 0.3$.

Strength of Instrument: Comparing Designs 8 with Design 1 with 10 features, we see that performance of the IV tree decreases and that of the causal tree remains almost the same when the instrument is weaker, so the performance gap of the two approaches decreases. However, the IV tree still has a better performance than the casual tree when $Corr(Z_i, T_i) = 0.4$.

Sample Size: From Table 3.2, we see that the coverage rate of the IV tree increases and mean-squared error decreases as sample size increases. These results are consistent with the theoretical results presented in Section 3.3.5. Note that the performance gap between the IV and causal trees does not become smaller as sample size increases. Fundamentally, this is because estimates from the causal tree are biased when the treatment is not randomly assigned, and such biases do not diminish as the sample size increases.

In summary, the results of the simulation studies suggest that the causal tree has poor coverage rates and large mean-squared errors when endogeneity is a concern. These results suggest that the estimated treatment effects from the causal tree are very different from the true treatment effects and confidence intervals of the estimated treatment effects may not cover the true treatment effects. Equipped with an exogenous instrument, the IV tree effectively corrects for these biases and significantly improves the coverage rate and reduces mean-squared errors of the estimates. The superior performance of the IV tree is consistent under different numbers of features, error distributions, model specifications, and feature types. The relative gap between the IV and causal trees decreases when there is less endogeneity or if the instrument is weak.

The performance of the IV tree increases as sample size increases whereas the coverage rate of the causal tree does not always improve with increasing sample size.

3.4.2 Comparison with the Generalized Random Forest

The generalized random forest (GRF) approach (Athey et al. 2019) can also be used to estimate heterogeneous treatment effects using IVs. The GRF method can be thought of as a "generalist" that can be used to estimate a broad range of parameters using moment conditions whereas the IV tree can be thought of as a "specialist" that is tailored for detecting heterogeneous treatment effects in observational studies. As such, the two approaches are complementary in the big data analytics tool kit. In this section, we conduct simulation studies to compare the accuracy and interpretability of these two approaches for analyzing heterogeneous treatment effects in observational data with unobservable features that affect both outcome variable and treatment assignment.

To allow comparison with the GRF method, we make some modifications to the IV tree described earlier. First, because directly comparing a tree to a forest is not fair, we grow an equal number of IV trees (which we refer to as an "IV Forest (IVF)") to compare with the GRF.[3] Second, following much of the existing literature on random forests (e.g., Breiman 2001), we use over-fitted instead of well-pruned trees to construct the IVF. Finally, we use the same local centering approach as in Athey et al. (2019, p21). That is, we first regress out the effect of the features on all the outcomes and then construct a forest using centered outcomes instead of original outcomes.

3.4.2.1 Synthetic Data Construction

We consider four designs to compare the performance of the two approaches. The first two designs are the same as those described in Section 3.4.1.1. The third and fourth designs are the same as those described in Athey et al. (2019). The details of each design are presented in Table 3.3. For each of the four designs, we consider instances with the number of features equal to 5, 10 and 20, with the sample size ranging from 1,000 to 5,000. Both the IVF and GRF are constructed using 100 trees and all features are used for splitting.[4] We compute mean-squared errors and split frequencies based on

[3] For a fair comparison, our forest is constructed using methods consistent with those in Athey et al. (2019). However, to formally extend a tree method to a random forest, more parameters need to be carefully evaluated and fine-tuned, which we leave for future research.

[4] As a default and in the simulation of Athey et al. (2019), the GRF considers all features for splitting when the number of features is less than or equal to 20 (see https://github.com/swager/grf for more details). Its performance remains almost the same or becomes worse when we restrict the number of features (e.g., to one-third or square root of the total number of features) for splitting.

a testing sample of size 5,000 and aggregate results based on 10 runs of the simulations.

3.4.2.2 Accuracy Comparison

Table 3.3 summarizes mean-squared errors of the IVF and GRF. We see that the IVF has smaller mean-squared errors in all scenarios except Designs 3 and 4 with five features and sample size of 1,000. The relative gap between the two approaches remains the same or increases as the sample size increases. When sample size equals 5,000, the relative gap ranges from 33% to 93%, depending on the scenarios. The primary reason the IVF generates more accurate estimates of treatment effects is that it uses the exact loss function for tree splitting whereas the GRF uses a gradient-based approximation.

Table 3.3: Mean-Squared Errors of IV Forest (IVF) and Generalized Random Forest (GRF)

		Sample Size											
#Features	Approach	1,000	3,000	5,000	1,000	3,000	5,000	1,000	3,000	5,000	1,000	3,000	5,000
		Design 1			Design 2			Design 3			Design 4		
5	IVF	0.18	0.10	0.06	0.02	0.01	0.00	0.43	0.15	0.11	0.49	0.21	0.13
	GRF	0.35	0.18	0.12	0.18	0.13	0.06	0.39	0.20	0.19	0.44	0.27	0.21
10	IVF	0.74	0.37	0.26	0.15	0.02	0.02	0.46	0.19	0.13	0.35	0.16	0.11
	GRF	1.28	0.64	0.42	0.28	0.17	0.12	0.50	0.31	0.22	0.55	0.35	0.26
20	IVF	2.85	1.90	1.48	0.58	0.28	0.21	0.48	0.18	0.15	0.39	0.19	0.12
	GRF	3.68	2.58	2.19	0.77	0.51	0.33	0.62	0.31	0.26	0.66	0.40	0.32

Note: Designs 1 and 2 have the form $Y_i = \sum_{k=1}^{3m} x_{ik} + \sum_{k=1}^{m} x_{ik}T_i + \varepsilon_i + \xi_i$, where $m = \{1, 2, 4\}$, and $x_{ik} \sim Norm(0, 1)$ in Design 1 and $x_{ik} \sim Bern(0.5)$ in Design 2. Designs 3 and 4 have the form $Y_i = \mu(X_i) + \tau(X_i)T_i + \xi_i$, where $\mu(X_i) = \sum_{k=1}^{2} max\{0, x_{ik}\}$, $\tau(X_i) = \sum_{k=3}^{4} max\{0, x_{ik}\}$ in Design 3 and $\mu(X_i) = max\{0, \sum_{k=1}^{2} x_{ik}\}$, $\tau(X_i) = max\{0, \sum_{k=3}^{4} x_{ik}\}$ in Design 4. All forests have 100 trees, and results are aggregated over 10 runs of the simulations.

3.4.2.3 Interpretability Comparison

In addition to estimating treatment effects, we also care about subject groupings. For example, in medical applications knowing which patients respond similarly to a treatment and which do not can provide clues to the underlying mechanism and thereby guide research into improved treatment alternatives. To evaluate the interpretability of the trees generated by the IVF and GRF approaches, we compare frequencies of splitting on both relevant and irrelevant features at each split depth, which is defined as the number of edges to the root node. A higher proportion of splits on relevant features implies greater interpretability. A shallower tree with a smaller number of subgroups is also easier to interpret.

Table 3.4: Split Frequencies of IV Forest (IVF) and Generalized Random Forest (GRF) in Design 2

		Sample Size											
Split on	Approach	1,000	3,000	5,000	1,000	3,000	5,000	1,000	3,000	5,000	1,000	3,000	5,000
		Depth 0			Depth 1			Depth 2			Depth 3		
Relevant Feature	IVF	979	1,000	1,000	10	0	0	2	0	0	0	0	0
	GRF	554	650	811	371	361	272	421	336	117	9	439	101
Irrelevant Features	IVF	21	0	0	756	687	685	231	372	314	0	91	74
	GRF	446	350	189	1,579	1,601	1,677	2,659	3,325	3,484	66	5,413	5,733

Note: Design 2 has the form $Y_i = \sum_{k=1}^{3} x_{ik} + x_{i3}T_i + \varepsilon_i + \xi_i$, where $x_{ik} \sim Bern(0.5)$. All forests have 100 trees, and results are aggregated over 10 runs of the simulations.

Table 3.4 summarizes split frequencies at the first four depths for Design 2, where x_3 is the only relevant feature and x_1, x_2 and x_4 are irrelevant features. At depth zero, the split frequency is 1,000 for 10 runs of the simulations for both approaches. We see that 979–1,000 trees in an IVF (compared with 554–811 trees in a GRF) split on x_3 at depth zero. At all depths, the GRF splits more on irrelevant features than does the IVF. Finally, most trees in the IVF have a depth of four or less whereas trees in the GRF are much deeper than those shown in Table 3.4. The combination of deeper trees and more splits at each depth leads to more subgroups with smaller sizes in GRF.

A main reason the GRF splits on irrelevant features is that it estimates parameters from moment conditions that require both the instrument and other exogenous features to be orthogonal to the error term. As a result, its splitting criterion is determined by both features that affect treatment effects and features that directly affect outcomes. In contrast, the objective of the IVF is to ensure the accuracy of treatment effect estimation. Its splitting criterion is determined (almost) exclusively by features that affect treatment effects.

Finally, an important reason the GRF has deep trees with small subgroups is that it takes into account only the mean of estimated treatment effects during splitting. It partitions observations into two child subgroups as long as the two subgroups have different average treatment effects. In contrast, the IVF considers both the mean and variance of estimated treatment effects (see first and second terms of Equation (3.8)) for tree splitting. It therefore balances the tradeoff between relevance (smaller subgroup size) and statistical reliability (less estimation noise).

3.5 Empirical Example: Teaching vs. Non-Teaching Hospitals

As mentioned in the Introduction, the IV tree proposed in this study can be applied to a wide array of applications, by defining a treatment to represent different things. In this section, we apply this approach to analyze the effect of teaching (vs. non-teaching) hospitals on outcomes for patients requiring colectomy. There has been a long-standing debate on whether teaching hospitals are better than non-teaching hospitals. While advocates for teaching hospitals argue that teaching hospitals are better because of the advanced technology, involvement in medical research, and treatment of rare diseases and complex patients, advocates for non-teaching hospitals argue that teaching hospitals are worse because of the substantial involvement of inexperienced residents and the attenuated role of senior physicians.

A number of studies in the medical literature have examined the outcome (e.g., mortality or complication rates) differences between teaching and non-teaching hospitals and found mixed results. For example, Masoomi et al. (2014), Thornlow et al. (2006), and Vartak et al. (2008) found that the difference between teaching and non-teaching hospitals is not statistically significant and teaching status is not a significant predictor of health care outcomes. Gopaldas et al. (2012) and Taylor et al. (1999) found that teaching hospitals are associated with lower complication rates and better survival. On the other hand, Duggirala et al. (2004), Fineberg et al. (2013), and Nandyala et al. (2014) found that teaching hospitals have higher post-operative complication rates than non-teaching hospitals.

Comparing teaching and non-teaching hospitals presents two major challenges. First, analyses based on observational data (e.g., medical records or administrative data) are subject to selection bias or unobserved confounding factors such as patients' preferences or adherence to treatment recommendations (Ayanian and Weissman 2002). Second, outcome differences between teaching and non-teaching hospitals may be heterogeneous across different procedures (Khuri et al. 2001) and patient subgroups. If teaching hospitals are better for some patient subgroups but worse (or similar) for other patient subgroups, studies focusing on the average outcome difference may not find significant results. The IV tree proposed in this study can address the these two challenges, because it corrects for potential endogeneity issues and partitions patients into subgroups with heterogeneous outcome differences.

To illustrate how the IV tree performs on real data, we focus on laparoscopic colectomy and use complication rate as the outcome measure to compare teaching and non-teaching hospitals. Laparoscopic colectomy is one type of colectomy that removes part or all of the large intestine (i.e, colon) to treat or prevent diseases of the colon including (a) inflammation of the digestive or gastrointestinal tract, (b) ulcers of the colon and rectum, and (c) malignant (cancerous) tumor in the colon or rectum.[5] This procedure is relatively complicated as it requires surgeons to pass a tiny video camera through one incision and special surgical tools through the other incisions. Around 22% of patients who underwent the procedure between 2005 and 2008 experienced at least one post-operative complication (Strasberg et al. 2011).

3.5.1 Data Description, Outcome Measure, and Feature Space

Our data consist of patient-level records for all inpatient discharges from all hospitals in New York State in 2011. They contain detailed hospital and patient information such as hospital identifiers, patient demographics and comorbidities, and principal and secondary diagnoses. The diagnosis codes allow us to identify surgery-related complications during hospitalization. We identify hospitals' teaching status through the web site of American Hospital Association (http://www.aha.org/).

Major complications of colectomy include wound infection, urinary tract infection, organ space infection, pneumonia, unplanned intubation, prolonged ileus/obstruction, bleeding, cardiac arrest, septic shock, systematic sepsis, myocardial infarction, renal insufficiency, deep venous thrombosis, pulmonary compromise, renal failure, cerebrovascular accident, and dehiscence. We focus on hospital-acquired instead of pre-existing complications to compare teaching with non-teaching hospitals. We are able to separate the two sources of complications, because the data indicate whether a complication was present on admission. The data do not allow us to track a patient's health status after he/she was discharged, so we focus on complications acquired during hospitalization.

In our data, 30.1% of patients had at least one of the 18 complications and 11.7% had two or more complications. Because a sizeable number of patients had more than one complication and different complications have different severity levels, we cannot simply use a binary variable to indicate whether a patient experienced at least one complication or simply count the total number of complications a patient experienced. To quantify both the number and severity of complications a patient experienced during his/her hospital stay, we convert complications into a numeric number that weights each complication by its severity.

The Accordion Severity Grading System is a scoring system that has been widely

[5] https://www.facs.org/media/files/education/patient.

used in the medical literature to systematically measure surgical complications by their severity levels (see e.g., Porembka et al. 2010, Strasberg et al. 2011). It stratifies complications into six grades, where Grade 1 and 2 complications are regarded as minor, Grade 3 as moderate, Grade 4 as serious and Grade 5 as life-threatening. Grade 6 complications refer to those that result in death of the patient and include death from any cause. These six grades of complications are associated with disutilities of 0.11, 0.26, 0.37, 0.60, 0.79, and 1.00 for Grades 1 to 6, respectively.

Strasberg et al. (2011) applied the Accordion Severity Grading System to analyze postoperative complications after laparoscopic colectomy and four other abdominal procedures. The severity score of a complication is calculated as the weighted sum of disutilities, where the weights are probabilities of the complication being classified into different grades. Table 4.2 summarizes relevant results from Strasberg et al. (2011). For the ease of discussion, we multiply all severity scores by 100. In our sample, the severity scores range from 11 to 100 with renal insufficiency being the least severe and death being the most severe complications. The incidence rates of the 18 complications range from 0.1% to 10.5% with cardiac arrest being the most rare and prolonged ileus/obstruction being the most common complications.

The features used to construct an IV tree include age group (below 50, 50-60, 60-70, 70-80, 80-90, and above 90), gender (male, female), race (white, black, Hispanic, Asian, Native American, and others), payer (Medicare, Medicaid, private insurance, self-pay, and others), location (urban, large rural, small rural, and isolated rural), income levels (from 1 to 4), and 29 different comorbidities.

3.5.2 Instrumental Variable Construction

We follow the approaches of KC and Terwiesch (2011) to construct a distance-based IV for teaching status. We first calculate the Euclidean distance between the centroid of patient i's zip code and that of hospital j, denoted as $Dist_{ij}$, using 5-digit zip codes included in the data. Though the actual travel distance is not the same as the Euclidean distance, existing studies have shown that the two distances are highly correlated with each other (Boscoe et al. 2012). We then estimate the probability of patient i going to hospital j, denoted as p_{ij}, using a multinomial logit model, $p_{ij} = exp(\delta Dist_{ij})/\sum_{j=1}^{J} exp(\delta Dist_{ij})$, where J indicates the total number of hospitals in patient i's choice set. Finally, we calculate the expected teaching status for each patient by summing up products of the probability of choosing a hospital and the hospital's teaching status over all hospitals: $\widehat{Teach}_i = \sum_{j=1}^{J} Teach_j \times p_{ij}$. We use $\widehat{Teach}_i$ as an

Table 3.5: Severity Score and Incidence Rate of Different Complications

Complication	Severity Score	Number of Cases	Rate of Incidence
bleeding	60	255	9.4%
cardiac arrest	26	4	0.1%
cerebrovascular accident	79	6	0.2%
death	100	25	0.9%
deep venous thrombosis	26	13	0.5%
dehiscence	44	32	1.2%
myocardial infraction	26	20	0.7%
organ space infection	35	33	1.2%
pneumonia	26	112	4.1%
prolonged ileu/obstruction	26	291	10.7%
pulmonary compromise	37	131	4.8%
renal failure	60	95	3.5%
renal insufficiency	11	50	1.8%
septic shock	66	15	0.6%
systematic sepsis	41	79	2.9%
unplanned intubation	79	89	3.3%
urinary tract infection	26	77	2.8%
wound infection	20	149	5.5%

Note: The severity score of a complication is calculated as the weighted sum of disutilities, where the weights are probabilities of the complication being classified into different grades (Strasberg et al. 2011). For the ease of discussion, we multiply all severity scores by 100.

IV for $Teach_i$.

Expected teaching status thus defined is a valid IV for our study, because it (1) correlates with the probability of choosing a teaching hospital and (2) does not correlate with unobservable sickness of a patient. That is, the closer a patient lives to a teaching hospital, the more likely the patient chooses a teaching hospital. However, how far a patient lives from a teaching hospital does not correlate with his/her sickness. The first condition is satisfied by construction. We follow existing studies (see e.g., Bartel et al. 2016, KC and Terwiesch 2012) to check the second condition by comparing observable sickness of patients with different levels of expected teaching status. We provide empirical evidence to support these two conditions in the Results and Discussion section.

In 2011, a total of 2,794 New York patients underwent elective laparoscopic colectomy surgeries at 156 hospitals. However, some of the patients do not have recorded zip code information, which is required to calculate the distance between a patient's home and a hospital in order to construct an IV for hospital teaching status. Exclusion of these patients results in a total of 2,723 patients discharged from 150 hospitals.

3.5.3 Results and Discussion

We check the strength of the IV by regressing teaching status over the instrument (i.e., expected teaching status) and other exogeneous variables for each subgroup of patients (i.e., the first stage of two-stage least square (2SLS) regression). The coefficients are significant at the 1% significance level in all cases. These results suggest that the IV has a strong first stage.

To check if patients living closer to a teaching hospital are sicker or healthier, we first stratify patients into three groups (with roughly one third of patients in each group) based on their distance to the nearest teaching hospital (Table 3.6) or their expected teaching status (Table 3.7) and then follow existing studies (see e.g., Bartel et al. 2016, KC and Terwisch 2012) to compare observable patient characteristics such as age, number of chronic conditions, and number of comorbidities. The results in Tables 3.6 and 3.7 suggest that travel distance and expected teaching status do not correlate with patient sickness.

We apply the IV tree and causal tree to the data and predict outcome differences between teaching and non-teaching hospitals for each patient. To provide an overall view of the difference between IV and causal trees in predicting outcome differences between teaching and non-teaching hospitals, we first categorize patients into four

Table 3.6: Relationship between Distance to Teaching Hospital and Patient Characteristics

Distance	Number of Patients	Patients' Mean Age	Number of Chronic Conditions	Number of Comorbidities
Short	932	63.5 (14.4)	4.4 (2.4)	2.1 (1.5)
Medium	883	64.4 (15.3)	4.5 (2.5)	2.0 (1.6)
Long	908	63.7 (14.5)	4.6 (2.7)	2.1 (1.7)
Total	2,723	63.8 (14.7)	4.5 (2.5)	2.1 (1.6)

Note: We stratify patients based on their distance to the nearest teaching hospital and analyze if patients living closer to a teaching hospital are sicker. Standard deviations are displayed in parentheses.

Table 3.7: Relationship between Expected Teaching and Patient Characteristics

Expected Teaching	Number of Patients	Patients' Mean Age	Number of Chronic Conditions	Number of Comorbidities
Low	908	63.8 (14.4)	4.7 (2.7)	2.1 (1.6)
Medium	912	63.9 (15.4)	4.6 (2.5)	2.1 (1.6)
High	903	63.8 (14.5)	4.3 (2.3)	2.1 (1.5)
Total	2,723	63.8 (14.7)	4.5 (2.5)	2.1 (1.6)

Note: We calculate expected teaching status for a patient by summing up products of his/her probability of choosing a hospital and the hospital's teaching status over all hospitals. We stratify patients based on their expected teaching status and analyze if patients with higher expected teaching status are sicker. Standard deviations are displayed in parentheses.

categories based on the statistical significance indicated by the IV tree and causal tree and then calculate the number of patients in each of the four categories (Table 3.8). We are most interested in Categories A and B, because patients in these groups highlight the difference between the two tree approaches. There are in total 703 patients (or 25.8%) in Categories A and B, suggesting that around one quarter of the patients could be misguided if we neglect the endogeneity issues and use the causal tree instead of the IV tree to compare hospitals. Table 3.9 compares the results of the IV and causal trees for 20 patients as examples.

Table 3.8: Comparison of IV Tree (IVT) and Causal Tree (CT) by Statistical Significance

			IVT			CT		
Category	Description	Cases	mean	min	max	mean	min	max
A	IVT significant, CT insignificant	455	9.51	7.71	10.99	6.57	−0.84	18.57
B	IVT insignificant, CT significant	248	8.5	3.36	13.65	10.94	6.11	19.41
C	IVT significant, CT significant	80	9.69	7.75	10.99	10.17	7.52	18.05
D	IVT insignificant, CT insignificant	1,940	6.69	3.36	13.65	5.96	−0.97	18.16
	Total	2,723						

Note: We apply the IV tree and causal tree to analyze the outcome differences between teaching and non-teaching hospitals across different patients. We categorize patients into four categories based on the statistical significance indicated by two trees and calculate the number of patients in each category.

To get a better sense of which patient subgroups benefit the most from non-teaching hospitals, we regroup patients by age, gender, and race. The average outcome differences between teaching and non-teaching hospitals for each subgroup are summarized in Figure 3.1. Generally speaking, male patients in their 90s benefit the most, followed by male patients in their 50s. Male patients in their 80s benefit the least. From Figure 3.1, we also see that that average outcome differences predicted by the causal tree are different from those predicted by the IV tree. For example, the causal tree predicts that female Asian patients in their 60s and 80s benefit the most and male Asian patients in their 50s benefit the least.

3.6 Conclusion

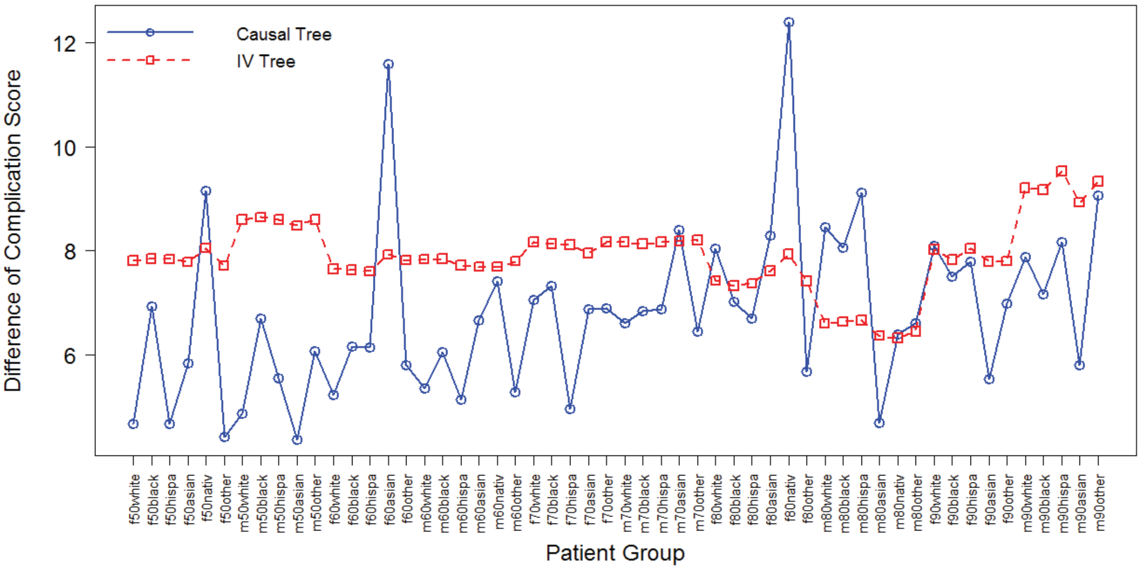
The big data revolution is driving a personalization revolution in diverse applications including customized marketing, precision medicine, and many others. Heterogeneous treatment effect analysis is a systematic and data-driven approach for personalization that stratifies individual subjects by their responses to a treatment. It offers organizations the potential to improve performance by personalizing their products and services

Table 3.9: Comparison of IV Tree (IVT) and Causal Tree (CT) for Different Patients

			IVT		CT	
Category	Patient	Patient Characteristics	mean	s.e.	mean	s.e.
A	1	50s, male, no comorb	7.71∗	6.00	8.31	6.84
	2	70s, female, 5 comorb	10.87∗	8.21	13.29	11.40
	3	70s, female, 5 comorb	9.50∗	6.14	7.75	9.00
	4	60s, female, 5 comorb	9.88∗	6.27	8.74	7.75
	5	90s, female, 3 comorb	9.51∗	6.11	8.29	7.55
B	6	60s, male, 2 comorb	10.59	8.75	9.55∗	7.35
	7	50s, female, 3 comorb	5.08	5.93	8.90∗	6.45
	8	80s, female, 3 comorb	9.15	7.23	9.05∗	6.81
	9	60s, male, 1 comorb	7.93	6.40	9.23∗	6.18
	10	70s, male, 1 comorb	7.71	8.59	9.94∗	7.66
C	11	80s, female, 2 comorb	9.51∗	6.11	9.92∗	6.86
	12	80s, female, 1 comorb	9.61∗	6.11	9.63∗	7.48
	13	60s, female, 2 comorb	9.99∗	6.24	9.92∗	7.13
	14	70s, female, 3 comorb	9.61∗	6.11	9.30∗	7.10
	15	70s, female, 0 comorb	9.61∗	6.11	10.00∗	7.47
D	16	60s, female, 4 comorb	11.51	12.28	9.31	7.49
	17	60s, male, 4 comorb	6.15	5.35	6.27	5.91
	18	60s, male, 3 comorb	6.60	7.31	6.48	8.93
	19	60s, male, 3 comorb	6.15	5.35	6.56	5.73
	20	40s, female, 2 comorb	3.50	5.30	5.92	5.73

Note: We compare the outcome differences between teaching and non-teaching hospitals estimated by the IV tree and causal tree for 20 sample patients. * $p < 0.1$.

Figure 3.1: Outcome Differences between Teaching and Non-Teaching Hospitals by Patient Subgroup



Note: Patients are regrouped by gender, age group, and race. We first estimate the outcome differences between teaching and non-teaching hospitals for each patient using the IV tree and causal tree, respectively. We then calculate the average outcome differences for a patient subgroup using all patients belonging to that group.

to individual customers.

We propose a new IV tree approach to address the challenges of using big observational data for heterogeneous treatment analysis. This approach combines the advantages of the recently developed causal tree and the classical IV method. It effectively partitions subjects into subgroups such that subjects in the same subgroup have similar treatment effects while those across different subgroups have different treatment effects. Our numerical studies demonstrate that the IV tree and IV forest methods are more accurate in estimating heterogeneous effects and more interpretable in understanding these effects than available alternatives.

We illustrate the use of IV tree by applying it to compare the outcomes of teaching and non-teaching hospitals for patients requiring laparoscopic colectomy. We find that outcome differences between teaching (designated as treatment) and nonteaching (designated as control) hospitals are heterogeneous across different patients, which is an important insight that links the debate about the roles of teaching and non-teaching hospitals with the conversation about personalized health care. This sample application also illustrates that the IV tree method can have a substantial impact on important treatment decisions.

CHAPTER 4

Personalized Health Care Outcome Analysis of Cardiovascular Surgical Procedures

4.1 Introduction

Choosing a health care provider for a major medical procedure can be literally a life or death decision. However, because they have historically lacked clear quality information about providers, most patients have made these important choices based on proximity or familiarity.[1] Even patients who have relied on physician referrals have been unable to rigorously evaluate their options, because the physicians themselves have also lacked objective data and therefore have had to rely on subjective reputation information.

Recognizing the critical need for more and better information about health care providers, government and private organizations have made various efforts to provide patient-oriented hospital ratings. For example, the Center for Medicare & Medicaid Services (CMS) maintains the Hospital Compare web site to compare Medicare-certified hospitals across the country and the US News provides aggregate hospital ratings for broad categories of procedures such as heart surgery and cancer. These, and other rating systems like them, compare hospitals based on risk-adjusted rates of mortality, complication and/or readmission, and assign scores or star ratings to hospitals based on their outcome measures.

However, a widely overlooked feature of these ratings is that they are based on population averages (hereinafter referred to as "population-average information"), which imply that the same hospitals are best for all patients. But this is an assumption built into population-average based ratings, rather than an empirical fact. To illustrate how

[1] http://www.infographicsarchive.com/health-and-safety/2014-healthgrades-american-hospital-quality-report-nation/

such ratings can be misleading, consider a simple example of three hospitals and two procedures — Coronary Artery Bypass Grafting (CABG) and Mitral Valve Surgery. Suppose the mortality rates of the three hospitals are 1%, 4% and 2%, respectively, for CABG patients, and 5%, 2% and 3% for mitral patients. If all three hospitals have a 50/50 mix of CABG and mitral patients, the overall mortality rates will be 3%, 3% and 2.5%, respectively. If hospitals are ranked according to overall mortality rate, then the third hospital will come out on top, even though it is not best for either procedure. Hence a population-average ranking on overall mortality rate will misguide patients (and their primary care physicians) in the choice of a hospital. By suggesting the same hospital for everyone, it will also contribute to a capacity imbalance.

In recognition that a hospital may perform well for some procedures and not as well for others, some states such as New York and Pennsylvania have begun publishing hospital quality report cards for individual cardiac surgeries such as coronary artery bypass grafting, aortic valve and mitral valve surgeries. Table 4.1 summarizes the risk-adjusted mortality rates and the relative ranking of six hospitals for three cardiovascular surgeries based on New York Cardiovascular Surgery Quality Report Cards 2011-2013.[2] The results show clearly that outcome differences are indeed heterogenous across procedures.

Table 4.1: Relative Performance of Hospitals for Different Procedures

Procedures		Lenox Hill Hospital	Mount Sinai	NYP-Columbia	NYP-Weill Cornell	Rochester General	St. Francis Hospital
Coronary Artery Bypass Grafting	Count	256	385	419	176	306	658
	Mortality	2.23%	1.80%	1.10%	1.74%	1.65%	1.54%
	Rank	6	5	1	4	3	2
Valve-Related Surgeries	Count	479	1820	2228	1303	1025	1831
	Mortality	3.30%	3.10%	2.88%	2.63%	4.91%	3.28%
	Rank	5	3	2	1	6	4
Percutaneous Coronary Intervention	Count	1551	4522	2541	1298	1569	2289
	Mortality	0.59%	0.92%	1.05%	1.50%	0.99%	0.82%
	Rank	1	3	5	6	4	2

Source: New York Cardiovascular Surgery Quality Report Cards 2011-2013.

But this still does not provide true patient-centric information, because patients requiring the same procedure differ in their demographics and severity of illness (Huckman and Kelly 2013). Hospital outcomes may be sensitive to these differences and the best hospital may be different for different patients.[3] To see whether outcome differ-

[2] https://www.health.ny.gov/statistics/diseases/cardiovascular/

[3] For example, diabetic patients in need of coronary bypass surgery have generally not been

ences are also heterogenous across other dimensions of patient characteristics, we need a way to group patients to generate statistically valid patient-centric outcomes.

Patient-centric ratings have obvious use in helping individual patients choose a hospital. But they have other important uses as well. The US government and private insurers are devoting considerable energy to designing payment structures that incentivize hospitals to improve quality. Most prominently, CMS has developed programs to link Medicare payments to hospital performance. For example, it launched the Readmission Reduction Program (RRP) in 2013 to penalize hospitals with excessive 30 day readmission rates and the Hospital Acquired Conditions Reduction Program (HACRP) in 2015 to penalize low performers with regard to hospital acquired infections.[4] In both programs, if a hospital's performance is below a threshold, the hospital is penalized for all its Diagnosis-Related Groups (DRGs). In 2015, more than 2,000 hospitals were penalized under RRP and more than 700 hospitals were penalized under HACRP.

A problem with both RRP and HACRP is that they rely on population-average data. As a result, they penalize some hospitals for all their procedures and do not penalize other hospitals for any procedure. As we noted above, low average performance does not necessarily mean that the hospital is poor at treating all patients, and high average performance does not necessarily mean a hospital provides good treatment to all patients. The result is a misalignment between the penalties (or lack of them) and hospital performance, and hence misalignment in the incentives to improve. Using patient-centric ratings allows payers such as CMS to assess hospital quality by patient group and thereby direct penalties more accurately at areas of poor performance.

In this paper, we examine six cardiovascular surgeries at thirty-five NY hospitals and address three key questions: (1) Are outcome differences between hospitals heterogenous across patient types? (2) If they are, how valuable is patient-centric information (that accounts for heterogeneity) to patients in selecting a provider? and (3) What impact would patient-centric information have on pay-for-performance systems in which providers reimbursements are based on patient outcome metrics?

Addressing these questions requires that we identify patient groups that exhibit

treated using the Bilateral Internal Thoracic Artery (BITA) grafting technique, because of concerns that they are at higher risk of infection involving the breast bone. However, the Cleveland Clinic found recently that BITA grafting can work very well for diabetic patients, except for those that are very overweight with diffuse atherosclerosis or widespread hardening of the arteries (see https://health.clevelandclinic.org/2014/11/the-best-bypass-surgery-option-for-diabetic-patients/ for more details). Similarly, surgeons at the Greenville Health System have found that patients with end stage renal disease (ESRD) require special care because they are at a higher risk for complications and death after surgical procedures including bypass grafting (Schneider et al. 2009).

[4] https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/index.html

significant differences in outcomes. A standard approach is to include interaction terms between provider indicators and patient characteristics as covariates in a multivariate regression model. This method works well when there is a small number of patient characteristics, but quickly breaks down when, as is the case here, the number of patient characteristics is large, because the combinations of characteristics increases exponentially with the number of patient characteristics.

An alternative approach is a nonparametric method that partitions patients into groups such that patients within the same group have similar outcome differences between providers. Unfortunately, while simple to state, it is not straightforward to find the best way to group patients. If we use all possible combinations of patient characteristics, we will have the same combinatorial problem as above. This will lead to groups that do not differ statistically and have sample sizes too small to yield good estimates of outcomes. But there is no natural way to divide patients into a priori groups.

Besides partitioning patients into heterogenous groups, we need to address potential endogeneity issues when using observational data to estimate outcome differences between providers. Unlike randomized controlled experiments where patients are randomly assigned to providers, patients in observational studies can choose providers based on available information. This non-randomized nature may create endogeneity issues if what drives patients' choice of providers is correlated with medical outcomes. We therefore use the instrumental variable tree approach, which combines the instrumental variable method with tree-based approach, to partition patients into groups and, at the same time, correct for potential endogeneity issues.

However, because the instrumental variable tree approach was designed to identify binary treatment effects, to identify heterogeneous provider effects when there are multiple providers, we must overcome three challenges. First, in addition to grouping patients, we also need to group providers because there may not be sufficient data to detect significant differences between all pairs of providers. Second, we need to derive from our groupings easy-to-understand outcome information for use by individual patients. Third, we need to find an instrument that correlates with the choice of providers but does not correlate with medical outcomes.

The above approach will enable us to evaluate differences in hospital outcomes and their heterogeneity across patient groups (Question 1). To address Question 2 about the value of patient-centric information in improving patient outcomes, we compare scenarios in which patients use patient-centric and population-average information to select the best provider for them. This enables us to characterize the magnitude of benefit to individual patients of having patient-centric, instead of population-average, data. Finally, we address Question 3 by using our estimates of patient outcomes under patient-centric and population-average information to examine how accounting for patient heterogeneity would impact Medicare pay-for-performance policies and the responses of hospitals to them.

4.2 Literature Review

There is growing interest in hospital quality from both the medical and operations management communities. The medical literature has focused primarily on identifying hospital characteristics that indicate better performance. For example, Keeler et al. (1992) compared 197 hospitals and found that teaching, large and urban hospitals are generally better than non-teaching, small, and rural hospitals for congestive heart failure, acute myocardial infraction, pneumonia, stroke or hip replacement. Birkmeyer et al. (2003), Gammie et al. (2009) and Vassileva et al. (2012) found high-volume hospitals tend to perform better than low-volume hospitals. Tsai et al. (2015) found that hospitals with boards that pay greater attention to clinical quality and use clinical quality metrics have more effective management practices and provide higher-quality care.

The operation management literature has taken a more detailed perspective by focusing on the impact of specific provider practices on performance. For example, Barro, Huckman and Kessler (2006), Clark and Huckman (2012), Huckman and Zinner (2008), and KC and Terwisch (2011) analyzed the impact of hospital specialization/focus on productivity and patient outcome; Clark, Huckman and Staats (2013), Huckman and Pisano (2006), KC and Staats (2012), KC et al. (2013) and Ramdas et al. (2017) analyzed the impact of related experiences on surgeon performance; Freeman et al. (2016), Jaeker and Tucker (2016) and Kim et al. (2014) analyzed the impact of workload on quality and patient outcome; Bavafa et al. (2018), Lu and Lu (2016), and Song et al. (2015) analyzed the impact of patient-physician communication, mandatory overtime laws and queue management on productivity and patient outcome.

A common assumption in both literatures is that the effects of quality driver are homogeneous across patient groups. Any study that gives a single ranking of providers or a single estimate of the impact of a practice on quality, regardless of patient group, is implicitly making this assumption. But a number of scholars have recognized the potential for this assumption to lead to inaccurate information to patients and have called for heterogeneous effect analysis in both patient care and quality assessment (see for example, FDA 2013, Gerteis 1993, IOM 2011, Kattan and Vickers 2004, Kent and Hayward 2007, Kravitz et al. 2004).

Existing models that incorporate heterogeneity usually assume latent classes of consumers with different tastes or that consumer tastes are random draws from a known distribution. For example, Xu et al. (2017) used a random coefficient multinomial logit model to characterize heterogeneous patient preferences in choosing doctors. Guajardo, Cohen and Netessine (2016) also used a random coefficient multinomial logit model to study the impact of service attributes on consumer demand in the US automobile industry. Lu et al. (2013) used a similar model to analyze how waiting in queue in the context of a retail store affects customers' purchasing behavior. While such modeling framework is useful in incorporating heterogeneous consumer preferences, they cannot systematically identify different combinations of characteristics that define heterogeneous consumer groups. As a result, it offers little practical guidance to individual consumers.

The machine learning literature, on the other hand, offers several useful frameworks to measure heterogeneity and to identify heterogeneous groups. For example, a few studies have proposed methods to analyze the heterogeneous treatment effects. Evaluating patient differences in the effect of a single treatment (e.g., a clinical trial of a new drug) is similar, although not identical, to evaluating patient differences in the relative outcomes across a set of providers. Hence, we discuss the literature on identifying heterogenous treatment effects as a guide to addressing heterogenous provider effects.

In two separate studies of biological markers in high-dimensional genomic data, Signovitch (2007) and Tian et al. (2014) applied the standard LASSO procedure with modified outcomes or covariates to determine from a large set of biological markers the subset of patients that can potentially benefit from a treatment. Imai and Ratkovic (2013) modified the standard LASSO procedure using different penalty factors for the covariates and treatment effects to distinguish the effect of treatment from that of covariates and to allow for the possibility of treatment effects with small magnitudes. Since they do not systematically partition patients into groups, these methods require users to define patient groups a priori. All of them apply a single global model to all observations, and assume that effects are linearly additive and errors follow some distribution.

Realizing that a single global model can not be applied to all observations, Zeileis, Hothorn and Hornik (2008) proposed to partition the observations into groups and apply separate local models such as linear regression or maximum-likelihood based models to individual groups. They proposed using a tree-based method to partition observations, where the feature with the highest instability is used to split groups, with a fluctuation test to analyze the parameter stability at a node. Su et al. (2009) modified the regression tree method to split the predictor space in a way that maximizes the square of the t-statistic for testing the null hypothesis that the average treatment effect is the same in the two potential groups. A tuning parameter is used to penalize complex trees with many terminal nodes, where the value of the parameter is determined through cross-validation based on the sum of squares of the split t-statistics. These methods split the predictor space based on model fit or a test-statistic, and do not use cross-validation to select the tuning parameter or to assess the goodness of fit of the estimated model. Furthermore, by their design these methods are better suited to outcome prediction than to heterogenous treatment effect analysis.

Recently, Athey and Imbens (2016) proposed a causal tree approach to analyze heterogeneous treatment effects in experimental studies where subjects are randomly assigned to treatment or control groups. We developed an instrumental variable tree approach that extends the causal tree approach to analyze heterogenous treatment effects in observational studies with endogeneity issues (see Chapter 3 for more details). The instrumental variable tree approach can be applied to analyze the heterogenous provider effect when there are two providers by interpreting one provider as the "treatment" group and the other provider as the "control" group. However, the instrumental variable tree approach cannot be used directly when there are multiple providers, because it is unclear which provider or providers should be designated as the treatment or control groups. Moreover, while the instrumental variable tree approach can be applied to each pair of providers, presenting such pairwise comparisons directly to patients is likely to be confusing since there may be hundreds of comparisons for a patient to process to come to a conclusion. In this study, we address these issues to derive easy-to-understand patient-centric information on a set of providers.

Our work also contributes to the recent stream of research to develop and apply machine learning techniques for better prediction or decision-making in operations management settings. For example, Ang et al. (2015) developed a new method that combines queueing theory and the LASSO procedure to improve the prediction of emergency department waiting time. Bertsimas et al. (2016) used several machine learning methods (LASSO, random forest and support vector machines) to predict the outcomes of clinical trials and optimize the test regimes. Bastani and Bayati (2016) developed a new efficient multi-armed bandit algorithm based on the LASSO estimator to tailor decision-making at individual levels. They illustrated the superior performance of this algorithm in warfarin dosing. Ban et al. (2016) introduced performance-based regularization to improve portfolio performance. Ferreira et al. (2015) used a regression tree approach to predict demand and to optimize price, which led to 9.7% revenue increase in a field experiment implemented at an online retailer.

4.3 The Empirical Model

In this section, we first describe the needs and challenges of generating patient-centric outcome information. We then introduce the instrumental variable tree approach from the machine learning literature and discuss how to extend it to identify heterogeneous outcome differences between providers across patient groups.

4.3.1 Problem Description

The basic problem in which we are interested is identifying the provider, or set of providers, with the highest likelihood of providing a good outcome for a given patient. The data available to us are the outcomes of prior patients at the various providers. However, because it is possible that outcomes are influenced by patient characteristics (e.g., age, comorbidities, etc.), prior patient outcomes are not equally relevant to the given patient. Patients with characteristics that match those of the given patient are more likely to be representative, than are patients with radically different characteristics. For instance, a 48-year old black woman with mitral valve disease and hypertension will probably get better information from outcomes of other middle aged mitral valve patients than she would from patients in their 90s with coronary artery disease.

While this insight is intuitive, it raises the important question of how similar a patient must be to provide useful information about likely outcomes. For example, are gender or race important? Or could a black female patient use outcomes from white male patients to help evaluate her options? Are only mitral valve patients relevant, or are patients with aortic valve disease also representative? Does hypertension matter? Or are outcomes from patients with other comorbidities, or no comorbidities, good indicators for our patient with hypertension? How much does age matter? Should our patient look only to outcomes for other 48 year olds, or should she consider patients within some wider window of ages? And so on. Ideally, a method for generating outcome information for a specific patient should also identify the cohort of patients from which this information should come.

The basic tradeoff involved in selecting a cohort is one of precision versus power. A very narrow cohort that closely matches the patient in question along all dimensions will be highly representative and hence precise in characterizing outcomes, but may be too small to offer statistical power needed to detect real and important differences between providers. A very broad cohort, which contains patients that may not resemble the patient in question, will be less precise in estimating outcomes but will have more power due to the larger sample size. The balance between precision and power should be struck endogenously by making use of the data itself.

A key characteristic of our problem is that we are seeking to characterize differences between provider outcomes. In contrast, most analyses focus on outcome prediction. The latter is relevant if a patient is choosing whether or not to receive a procedure. For example, to decide whether the risk of heart surgery is justified by the benefits, we need an estimate of the mortality rate from the procedure. However, once we have decided to receive a procedure and must decide on a provider, it is the difference in the mortality rates between the candidate providers that matters. In a deterministic world, where we know the absolute mortality rates, we can compute the differences via simple subtraction. But in a statistical world, where we can only estimate the rates, a method that focuses on prediction of the absolute rates may not yield the most accurate estimate of the differences between rates, because the factors that affect outcomes may not be different from those affect outcome differences. We focus explicitly on estimating differences between providers, in the following discussion of instrumental variable tree approach, and in the subsequent empirical analysis.

Finally, because providers have different mixes of patients, we need to control for patient demographics, common comorbidites and other patient characteristics that affect patient outcomes. While it is straightforward to control for observable features, there are often unobservable features that affect both provider choice and patient outcomes. For example, health conscious patients may be more likely to choose high-performing providers for a better treatment. They are also more likely to receive better outcomes due to their healthier living styles. But whether a patient is health conscious or not is not observable to us as researchers. Endogeneity issues like this will create biases in estimating the outcome differences between providers.

4.3.2 Instrumental Variable Tree Approach

We use the instrumental variable tree approach developed in Chapter 3 to analyze heterogenous differences between providers for several reasons. First, this approach recursively partitions patients into heterogenous groups such that patients in the same group exhibit similar outcome differences across providers and those in different groups have different outcome differences. Second, the instrumental variable tree approach incorporates the instrumental variable method to correct for potential endogeneity issues, so we can obtain unbiased estimates of outcome differences for different patient groups. Third, this semiparametric approach does not assume that effects of different features are linearly additive and allows features to interact in highly nonlinear and complex ways. Finally, we can use techniques such as cross-validation to compare different tree models and select the one that best balances the tradeoff between prediction error and model complexity.

The instrumental variable tree approach builds on ideas developed by Athey and Imben (2016), who proposed the causal tree approach to analyze heterogenous treatment effects in experimental studies (e.g., clinical trial) where subjects are randomly assigned to treatment or control groups. Because the treatment and control groups have the same mix of subjects, the causal tree approach estimates the treatment effect $D(x_{l_j})$ of group l_j with feature x_{l_j} using the difference between average outcomes of the treatment and control groups (denoted as $\overline{y}_{1l_j}$ and $\overline{y}_{2l_j}$). That is

$$\begin{aligned}\hat{D}_{CT}(x_{l_j}) &= \overline{y}_{1l_j} - \overline{y}_{2l_j} \\ &= \tfrac{1}{N_{1l_j}} \textstyle\sum_{i \in l_j, T_{1i}=1} Y_i - \tfrac{1}{N_{2l_j}} \textstyle\sum_{i \in l_j, T_{1i}=0} Y_i\end{aligned}$$

where N_{1l_j} and N_{2l_j} denote the numbers of subjects in group l_j that receive the treatment and control, respectively, and T_{1i} indicates whether subject i receives the treatment.

When subjects are not randomly assigned to treatment or control groups, $\hat{D}_{CT}(x_{l_j})$ leads to a biased estimate of $D(x_{l_j})$ and the causal tree constructed using the biased estimates of treatment effects partitions subjects into wrong groups. To address this issue, we (see Chapter 3) developed the instrumental variable tree approach. The instrumental variable tree approach corrects for potential endogeneity issues and thus obtains unbiased estimates of treatment effects. Below, we first describe the instrumental variable tree approach and discuss how an analogous approach can be used to identify heterogenous provider effects when there are two providers, and then extend the instrumental variable tree approach to identify heterogenous provider effects when there are multiple providers.

4.3.2.1 IV Tree with Two Providers

The instrumental variable tree approach was originally developed to identify binary treatment effects in observational studies where the treatment is not randomly assigned. When there are two providers, we can designate one provider as the treatment group and the other provider as the control group. The treatment effect estimated using the instrumental variable tree approach can be interpreted as the provider effect in this study. To describe the instrumental variable tree approach, we let $T_{1i} = 1$ indicate whether patient i receives the treatment from provider 1 (instead of provider 2), S_i denote unobservable features (e.g., echocardiogram) that affect both patient outcomes and the choice of a provider, and ξ_i be the idiosyncratic error. For patient i in group l_j, the potential outcome Y_i can be written as

$$Y_i \quad = \quad \alpha_{0l_j} + \alpha_{1l_j} T_{1i} + \epsilon_i, \forall i \in l_j$$

where $\epsilon_i = \alpha_{2l_j} S_i + \xi_i$.

The parameter of interest is α_{1l_j}, which describes the average provider effect for group l_j. Let Z_i be an instrumental variable (e.g., proximity to a provider) that (1) correlates with the choice of a provider T_{1i} (i.e., satisfying the relevance condition), and (2) does not correlate with the error term ϵ_i (i.e., satisfying the exogeneity condition). The instrumental variable tree approach estimates provider effect α_{1l_j} using

$$\hat{D}_{IV}(x_{l_j}) \quad = \quad \tfrac{Cov(Y_i, Z_i)}{Cov(T_{1i}, Z_i)}, \forall i \in l_j$$

and the variance of $\hat{D}_{IV}(x_{l_j})$ using

$$Var[\hat{D}_{IV}(x_{l_j})] \quad = \quad \tfrac{Var(\epsilon_i)}{N_{l_j} Var(T_{1i})[Cor(T_{li}, Z_i)]^2}, \forall i \in l_j$$

where N_{l_j} denotes the number of subjects in group l_j. To estimate $Var(\epsilon_i)$, the instrumental variable tree approach uses the residuals $\hat{\varepsilon}_i = Y_i - \hat{\alpha}_{0l_j} - \hat{\alpha}_{1l_j} T_{1i}$, where $\hat{\alpha}_{0l_j}$ and $\hat{\alpha}_{1l_j}$ are the instrumental variable estimates. A consistent estimator of $Var(\epsilon_i)$ is $(\sum_{i \in l_j} \hat{\varepsilon}_i^2)/(N_{l_j} - 2)$, where $N_{l_j} - 2$ is used for the degrees of freedom correction.

To prevent the model from identifying spurious correlation between the features and outcomes as treatment effects, the instrumental variable tree approach splits data into two parts — one part for training (S^{tr}) and the other part for estimation (S^{es}). The splitting criteria of the instrumental variable tree approach was derived by minimizing the expected mean-squared error (denoted as $EMSE^{IV}$) over testing (S^{te}) and estimation samples

$$\begin{aligned} EMSE^{IV}(S^{te}, S^{es}) &= E_{S^{te},S^{es}} MSE(S^{te}, S^{es}) \\ &= -E_{S^{te}}[\hat{D}^{tr}_{IV}(X_i)^2] + E_{S^{te},S^{es}}[Var(\hat{D}^{es}_{IV}(X_i))] \end{aligned}$$

where $\hat{D}^{tr}_{IV}(X_i)$ and $\hat{D}^{es}_{IV}(X_i)$ denote the estimates of provider effects for patient $X_i \in S^{te}$ based on the training and estimation samples, respectively.

The instrumental tree approach starts at the top of the tree, which consists of a single group called the "parent group", and successively makes binary splits of groups based on the feature that reduces $EMSE^{IV}(S^{te}, S^{es})$ the most. This recursive partitioning process leads to a very large initial tree, which is then pruned recursively based on the weakest links to obtain a series of subtrees. Similar to the classification and regression tree approaches, the instrumental variable tree approach uses a tuning parameter α to balance expected mean-squared error and tree complexity, $EMSE^{IV}(S^{te}, S^{es}) + \alpha M$, where M denotes the number of terminal nodes of a subtree, and uses cross-validation to choose the subtree that minimizes average mean squared errors. Finally, the outcome difference for each terminal node is estimated using an independent estimation sample using the instrumental variable method.

4.3.2.2 IV Tree with Multiple Providers

While it is straightforward to apply the instrumental variable tree approach to analyze heterogeneous provider effects for two providers, we need to clear several hurdles to extend the method to multiple providers. Recall that the instrumental variable tree approach splits on features in a way that minimizes $EMSE^{IV}(S^{te}, S^{es})$. When there are multiple providers, it is unclear which provider or set of providers should be considered as the treatment group and which as the control group. This implies that we must partition providers, as well as patient groups. Note that partitions of providers can be different for different patient groups and vice versa.

There are several options for addressing this issue. Some of these require predefined provider groups, while others involve modifications of the splitting criteria of the instrumental variable tree approach to accommodate differences of all pairs of providers. For instance, the instrumental tree approach can be applied directly if a provider itself is considered as a group and all the other providers are considered as another group. We can build the instrumental variable tree using patient characteristics and the provider indicator as features. If the tree splits on the provider indicator, it indicates that the provider differs from the other providers as a group. We can estimate outcome difference between the provider and the other providers using the procedures discussed earlier. The problem with this approach, however, is that the derived outcome information can be confusing, because the baseline group changes as we move to compare another provider with its peers. As a result, a patient cannot directly compare the outcomes of two providers when his/her choices of providers are limited.

An alternative is to modify the splitting criteria. For instance, we can partition patients into groups such that, within each group, there is a large outcome variation across all providers. To do this, the splitting criteria can be modified to $-E_{S^{te}}[\widetilde{D}^{tr}_{IV}(X_i)^2] + E_{S^{te},S^{es}}[Var(\widetilde{D}^{es}_{IV}(X_i))]$, where $\widetilde{D}_{IV}(X_i)$ captures the average outcome differences between all pairs of providers for patient i. That is, $\widetilde{D}_{IV}(X_i) = \frac{1}{N}[\sum_{j \neq k} |\hat{D}^{IV}_{jk}(X_i)|]$, where N denotes the number of unique pairs of providers. The problem with this approach, however, is that the groups differentiating one pair of providers may be different from those differentiating another pair of providers. Consider a simple example where Provider 1 is better than Provider 2 only for young patients and Provider 3 is better than Provider 4 only for male patients. The instrumental variable tree approach with above modified objective function is not suitable because it will result in a universal partition that is homogeneous across all provider pairs, and hence is not sensitive to the heterogeneous differences across provider pairs.

We address these issues by applying the instrumental variable tree approach to each pair of providers. While the approach is methodologically sound, it poses significant interpretation difficulties. For example, a patient considering 10 providers would have to examine 45 pairwise comparisons, which is likely to lead to confusion. To avoid this, we develop a two-stage approach. In the first stage, we analyze pairwise provider differences. In the second stage, we condense the results into a form that enables a patient to make direct comparisons between any provider and the state average. More specifically, we first estimate the outcome difference between a provider j and any of the other providers. To do this, we build $N-1$ instrumental variable trees using provider j and the other $N-1$ providers one at a time. From these trees, we can estimate the outcome differences between providers j and k for patient i, $\hat{D}^{IV}_{jk}(X_i), \forall j \neq k$. We then use the estimated results to derive patient-centric outcome information based on the outcome difference between each provider and the state average. To formalize this, we let $D_{j,SA}(X_i)$ denote the difference between provider j and the state average of H

providers,

$$\begin{aligned} D_{j,SA}(X_i) &= E[Y_j(X_i) - \tfrac{1}{H}(Y_1(X_i) + Y_2(X_i) + ... + Y_H(X_i))] \\ &= \tfrac{1}{H}\textstyle\sum_{k\neq j} E[Y_j(X_i) - Y_k(X_i))] \\ &= \tfrac{1}{H}\textstyle\sum_{k\neq j} D_{jk}(X_i) \end{aligned}$$

Intuitively, we can estimate $D_{j,SA}(X_i)$ using $\frac{1}{H}\sum_{k\neq j}\hat{D}^{IV}_{jk}(X_i)$. Note that, because we partition patients into groups based on the outcome differences between two providers, the groups we identify by comparing providers j and k may be different from those identified by comparing providers j and l. For example, if provider j is better than provider k at treating male patients but better than provider l at treating white patients, the causal trees will partition patients into {male, female} when comparing providers j and k and {white, non-white} when comparing providers j and l. However, this does not affect our estimation of outcome differences between provider j and the state average.

4.4 Empirical Setting and Data

We choose cardiovascular diseases (commonly known as heart diseases) as the empirical setting for personalized health care outcome analysis for several reasons. First, cardiovascular diseases are the leading cause of death worldwide (WHO 2011). Each year, about 17.5 million people die from cardiovascular diseases, which accounts for one in every four deaths, and this number is expected to grow to more than 23.6 million by 2030.[5] Second, cardiovascular surgeries are relatively complicated procedures. They require sophisticated skills, advanced technology and intensive post-surgical care, which makes them candidates for sizable variations across providers (hospitals or surgeons). Third, cardiovascular surgeries include several different types of procedures, each requiring a different set of skills and technology. As a result, a hospital may perform well for some procedures but not as well for others.

Cardiovascular diseases refer to (a) conditions when the blood vessels are narrowed or blocked, which can lead to heart attack, (b) chest pain or stroke, and (c) conditions that affect the heart's muscles, valves or rhythm. Cardiovascular surgeries are operations performed by surgeons on the heart and blood vessels to repair the damage caused

[5] https://www.heart.org/idc/groups/ahamah-public

by diseases or disorders of the cardiovascular system. In this study, we focus specifically on three cardiac surgeries — Mitral Valve Replacement (MVR), Aortic Valve Replacement (AVR) and Coronary Artery Bypass Grafting (CABG), and three vascular surgeries — Abdominal Aortic Aneurysm (AAA) repair, Carotid Endarterectomy (CE) and Lower Extremity bypass Graft (LEG).

4.4.1 Data Description and Preparation

Our study makes use of data from New York State that consist of patient-level records of all in- and out-patient discharges from all hospitals in New York from 2008–2012. The data contain detailed clinical and resource use information, including admission status, patient demographics and comorbidities, hospital identifiers, and principal and secondary diagnoses. For each discharge, the data indicate the type of surgery a patient underwent. They also record whether a patient experienced any complications during the procedure or post-surgery hospitalization.

We identify discharges related to the six cardiovascular procedures under this study by using related clinical codes in the International Classification of Disease (9th revision). From 2008–2012, a total of 124,895 patients with cardiovascular diseases were discharged from 144 hospitals. Because some of the hospitals did not perform cardiovascular surgeries every year or had a low volume, we restrict attention to the 41 cardiac hospitals compared by the New York State of Health for Cardiovascular Surgery Quality Report Cards. However, six of these hospitals did not perform vascular surgeries, so we further narrow our focus to the other 35 hospitals that perform all the six cardiovascular surgeries discussed earlier. This results in a total of 107,252 discharges over the five year period. We focus on isolated surgeries and exclude patients who underwent multiple types of surgeries (6,950 of the sample). This allows us to characterize patient outcomes at each hospital for each surgery type. In addition, we exclude patients with missing information such as admission status. Our final sample contains a total of 99,378 discharges.

4.4.2 Outcome Measure and Feature Space

We focus on hospital acquired complications to characterize surgical outcomes, because they capture a wide range of negative patient experiences and show substantial variation across hospitals. But outcome differences between providers can be evaluated in terms of other metrics such as readmission, mortality, or a composite score that combines them, without changing the overall conclusions of this study. We identify complications using the diagnosis codes provided in the data and focus on hospital acquired conditions rather than pre-existing conditions. We are able to separate the two types of complications because the data indicate whether each diagnosis was present at admission. We focus on 23 cardiovascular surgery related complications[6] and use them collectively as an outcome measure (STS 2016, van Tuinen et al. 2005, Williams et al. 1965).

In our sample, 29.58% patients had at least one of the 23 complications, while 10.55% had two or more complications. Because a sizeable number of patients had more than one complications, we cannot simply use a binary variable to indicate whether a patient experienced at least one complication. In addition, the 23 complications have different severity levels. For example, complications such as pulmonary embolism or insufficiency are relatively easy to cure, while complications such as coma and multi-organ failure are likely to lead to patient deaths (Glance et al. 2007, Reddy et al. 2013). Therefore, we cannot simply count the number of complications a patient experienced. To capture both the number and the severity of complications associated with a patient during the surgery and hospital stay, we need to translate complications into a numeric score that weights each complication by its severity.

To do this, we adapt the approach inspired by Elixhauser et al. (1998). The Elixhauser comorbidity index is a vector of 30 binary variables in which each represents the existence of a comorbidity. To describe the overall sickness of a patient and to weight the severity of individual comorbidities, van Walraven et al. (2009) modified the Elixhauser comorbidity index into a single numeric Elixhauser comorbidity score by using a backward stepwise multivariate logistic regression to determine the correlation between each comorbidity and in-hospital mortality. The parameter estimates of the regression model were translated into a vector of weights by Sullivan et al. (2006). The Elixhauser comorbidity score, which is calculated as the dot product of the index vector and the vector of weights, has been widely used in medical research studies (Kang et al. 2010, Menendez et al. 2014, Silverstein et al. 2008). In this paper, we use the same approach to develop a complication score for cardiovascular surgical outcomes. The complications and their weights are summarized in Table 4.2. The average complication score for each procedure in our study ranges from 0.11 (for CE) to 1.65 (for AAA) and the average across all procedures is 0.68 (Table 4.3).

[6] The complications are stroke, aortic dissection, renal failure, ventilation, multi-organ failure, coma, cardiac arrest, sepsis, gastrointestinal events, tracheal reintubation, surgical complications, tamponade, wound infection, renal dialysis, mediastinum, reoperation for bleeding, pneumonia, pulmonary embolism, heart block, myocardial infarction, pulmonary insufficiency, surgical E codes and other cardiac complications.

Table 4.2: Weights of Different Complications

Complication	Coefficient	Std. Err.	Weight
Aortic Dissection	$3.16***$	0.33	7
Coma	$2.76***$	0.25	6
Multi-Organ Failure	$2.16***$	0.07	5
Cardiac Arrest	$1.79***$	0.09	4
Renal failure	$1.46***$	0.05	3
Tracheal Reintubation	$1.22***$	0.06	3
Sepsis	$1.03***$	0.14	2
Stroke	$1.03***$	0.12	2
Surgical Complication	$1.11***$	0.15	2
Tamponade	$1.02***$	0.14	2
Ventilation	$0.85***$	0.07	2
Gastrointestinal Event	$0.44***$	0.10	1
Pulmonary Insufficiency	$0.46***$	0.06	1
Constant	$-4.93***$	0.04	

Note: Robust standard errors are clustered by hospital. The outcome variable is death during hospitalization. Complications dropped from backward stepwise multivariate logistic regression based on statistical significance include wound infection, renal dialysis, mediastinum, reoperation for bleeding, pneumonia, pulmonary embolism, heart block, myocardial infarction, surgical E codes and other cardiac complications.
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 4.3: Summary of Complication Score for Different Procedures

Procedure Name	Count	Mean	Std. Dev.
Aortic Valve Replacement	20,061	0.99	2.30
Coronary Artery Bypass Graft	46,098	0.66	1.80
Mitral Valve Replacement	5,097	1.47	2.80
Abdominal Aortic Aneurysm	1,356	1.65	2.86
Carotid Endarterectomy	14,539	0.11	0.77
Lower Extremity Bypass Graft	12,227	0.41	1.47
Total	99,378	0.68	1.90

The features we use to construct the instrumental variable trees include six cardiovascular procedures (CE, CABG, LBG, AAA, AVR and MVR), patient genders, races (white, black, Hispanic, Asian, Native American, and others), admission statuses (emergent, urgent and elective), six age groups (below 50, 50–60, 60–70, 70–80, 80–90 and above 90) and five major comorbidities (chronic heart failure, chronic lung disease, diabetes, hypertension and renal failure) of cardiovascular diseases (STS 2016). Considering all these features results in a total of 6 procedures × 2 genders × 6 races × 3 admissions × 6 ages × 2^5 comorbidities = 41,472 different combinations of patient features.

4.4.3 Instrumental Variable Method

To correct for potential endogeneity issues, we apply the instrumental variable tree approach and use distance to construct an instrument for the provider indicator. For patient i and providers j and k, we first calculate the Euclidean distances d_{ij} and d_{ik} using 5-digit zip codes of patients' home and hospital addresses. We then compare the two distances to determine if provider j is closer than provider k to patient i. Then the indicator function $\mathbb{1}[d_{ij} < d_{ik}]$ can be used as an instrument for the provider indicator T_{ji}.

Prior studies have used distance or a function of distance as an instrumental variable to compare providers (McClellan et al. 1994,Brooks et al. 2006, KC and Terwiesch 2011). As in these studies, distance is an appropriate instrument for our purpose because (1) the distance between a patient and a provider affects the choice of the provider, and (2) how far a patient lives from a provider does not correlate with the sickness of the patient. We provide empirical evidence supporting these two criteria in Appendix C.1.

4.5 Results and Discussion

As described in Section 4.3, to identify hospitals that are statistically significantly different from the state average for certain patient groups, we first construct instrumental variables trees for each pair of hospitals, which requires a total of $35 \times 34/2 = 595$ trees. For each patient, we estimate the differences in complication score between a hospital and the state average, and estimate the standard errors of the differences using the bootstrap method.

Table 4.4 summarizes the results for six example patients each described by a combination of procedure type, age and comorbidities. The best hospital for each patient is highlighted in bold. We observe that, while some hospitals (e.g., Hospital 1) are uniformly better than the state average for all six patients, others (e.g., Hospital 35) are worse than the state average for majority of the patients. However, for hospitals that are uniformly better (or worse) than the state average, the magnitude of the differences varies for individual patients. For example, Hospital 1 is better than the state average by 0.24 for the 1st patient (CE, 70s, 4 comorbidities) and by 0.84 for the 5th patient (AAA, 70s, no comorbidity). There are also hospitals that are better than the state average for some patients but worse for others. For example, Hospital 12 is better for the 1st (CE, 70s, 4 comorbidities), 2nd (CABG, 70s, 1 comorbidity) and 5th (AAA, 70s, no comorbidity) patients but worse for the 3rd (AVR, 40s, 2 comorbidities), 4th (MVR, 50s, 4 comorbidities) and 6th (LBG, 70s, 4 comorbidities) patients. These results indicate that outcome differences between hospitals are indeed heterogenous across patients, and that different patients have different sets of hospitals that are significantly better that the state average.

Of course, Table 4.4 only shows six patients as examples. We have analyzed the outcome differences across hospitals for all of the patients in this study. To provide a visual illustration of the heterogeneity in outcomes across hospitals for different patients, we group patients by procedure type, age group and comorbidities, which are the most important features affecting outcome differences.[7]

For each patient group, we use $Y_{ijk} \in \{-1, 0, 1\}$ to indicate whether hospital j is statistically significantly worse than, the same as, or better than the state average at a 10% significance level for patient i in group k. Then we calculate the overall performance of hospital j for patient group k using $\bar{Y}_{jk} = \frac{1}{N_{jk}} \sum_{i=1}^{N_{jk}} Y_{ijk}$ and present the results in a heat map (Figure 4.1), where the yellow/red colors indicate that a hospital's overall performance is better/worse than the state average, and the intensity of the colors indicates the fraction of patients in a cell for which a hospital is better/worse than the state average.

From Figure 4.1, we observe that many of the cells in the middle (i.e., those associated with hospitals 13–30) are orange, which indicates that these hospitals are not significantly different from the state average for many patient groups. The majority of the cells in rows at the top (e.g., those associated with hospitals 1–3) have the color of yellow, indicating that these hospitals are better than the state average for most

[7] Note that the actual grouping of patients is determined by the instrumental variable tree approach. Because it is impossible to summarize all results from the trees in a single figure or table, we regroup patients based on the most important features to illustrate the heterogeneity in outcomes across hospitals for different patients.

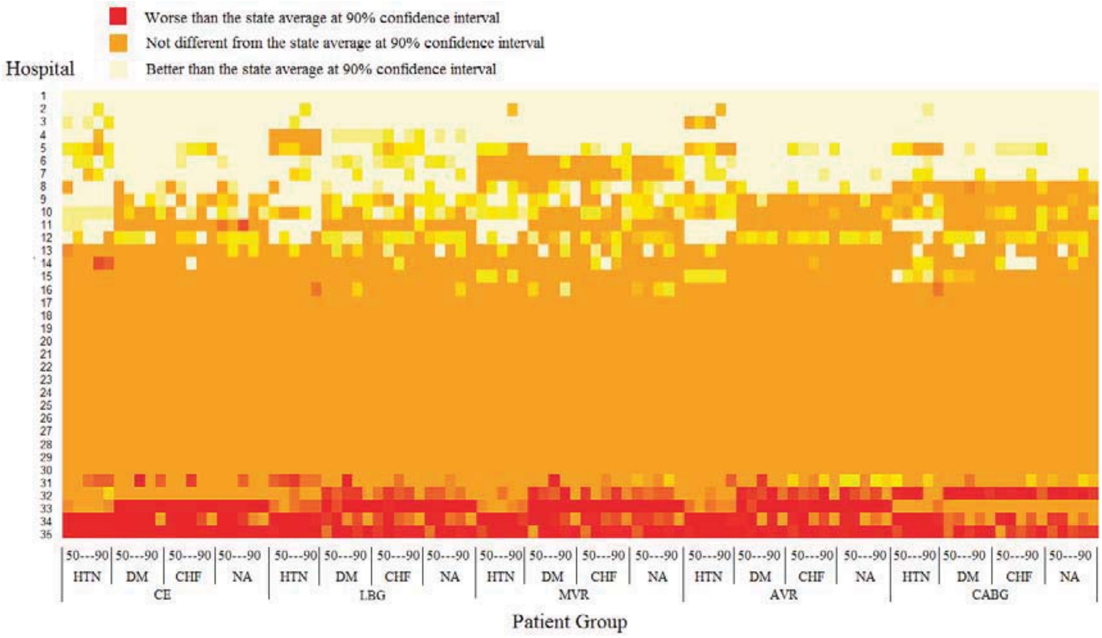
Table 4.4: Comparison of Complication Score with the State Average for Different Patients

Hospital	CE, 70s	CABG, 70s	AVR, 40s	MVR, 50s	AAA, 70s	LBG, 70s
Index	4 Comorb	1 Comorb	2 Comorb	4 Comorb	0 Comorb	4 Comorb
	(1)	(2)	(3)	(4)	(5)	(6)
1	−0.24+++	−0.27+++	**−0.62**+++	**−0.59**+++	**−0.84**+++	**−0.52**+++
2	−0.14	−0.22+++	−0.26++	−0.24++	−0.40++	−0.26+++
3	−0.46+	**−0.42**+++	−0.30	−0.34++	−0.73+++	−0.28++
4	−0.19	−0.24+	−0.10	−0.12	−0.49+++	0.02
5	**−0.75**+++	0.01	−0.27+++	−0.12	−0.38+++	−0.15
6	−0.56+++	−0.07	−0.28+++	−0.20++	−0.35+++	−0.14
7	−0.45++	−0.06	−0.42+++	−0.10	−0.35++	−0.24+
8	−0.45+++	0.00	0.31-	0.16	−0.43+++	−0.16++
9	−0.25+++	−0.14++	−0.34+++	−0.34+++	−0.13	−0.35+++
10	−0.11+	−0.12+	−0.33+	−0.33+++	−0.09	−0.30+++
11	−0.40++	−0.29+++	−0.28++	−0.23++	0.16	−0.29++
12	−0.15++	−0.13+	0.36-	0.33-	−0.28++	0.31-
13	−0.12	−0.02	−0.21++	−0.20++	−0.27++	0.05
14	−0.09	−0.10	−0.15	−0.05	0.00	−0.07
15	−0.29++	−0.10+	0.02	0.13	0.10	0.16-
16	0.09	0.15	−0.13	−0.13	−0.21+	−0.01
17	−0.31	−0.28	−0.53	−0.54	−0.72	−0.48
18	0.11	0.11	0.03	0.18	0.10	0.16
19	0.23	0.00	−0.12	−0.17	0.08	−0.16
20	0.56	−0.11	0.19	−0.13	0.10	0.02
21	0.80	1.19	0.61	0.58	1.43	0.93
22	1.16	−0.07	0.05	0.05	0.36	0.05
23	0.79	0.52	0.50	0.58	0.37	−0.10
24	−0.07	0.23	0.27	0.27	0.29	0.37
25	0.30	−0.27	0.25	0.03	0.00	−0.07
26	0.81	0.67	0.59	0.47	0.69	0.49
27	−0.02	−0.07	−0.05	−0.09	0.37	0.06
28	0.35	0.80	0.53	0.54	0.36	0.49
29	0.37	0.11	0.13	0.02	0.47	0.01
30	−0.12	−0.18	−0.09	−0.11	0.36	0.12
31	0.03	−0.11+	0.05	0.16	−0.04	0.22-
32	−0.34+++	0.09	−0.27+++	−0.23++	0.40--	−0.14+
33	0.04	0.10	0.14	0.19	0.59--	0.10
34	0.11	0.38---	0.66---	0.65---	0.10	0.54---
35	0.97---	0.17-	0.20-	0.20-	0.18	0.23--

+++, ++, +: better than state average at 99%, 95% and 90% confidence level

---, --, -: worse than state average at 99%, 95% and 90% confidence level patient groups. In contrast, the red color of the cells in rows at the bottom (e.g., those associated with hospitals 33–35) indicates that these hospitals are worse than the state average for most patient groups. Rows near the top having colors of yellow and orange indicate that the corresponding hospitals are better for some patient groups, but are not statistically different from the state average for other patient groups. Likewise, rows near the bottom with a mixture of red and orange cells indicate that these hospitals are worse for some patient groups but are not significantly different from the state average for other groups. Interestingly, there are hospitals (e.g., 11, 14, 16 and 31) that are significantly better than the state average for some patient groups but are significantly worse than the state average for other patient groups. Hence, the answer to Question 1 in the Introduction is yes; outcome differences between hospitals are heterogenous across patient types.

Figure 4.1: Comparison of Complication Score for Different Patient Groups (IV Tree)



Note: Patients are grouped by age group (i.e., 50s to 90s), comorbidity and surgery. Acronyms for comorbidities: HTN - hypertension, DM - diabetes, CHF - chronic heart failure, NA - no comorbidities. Acronyms for surgeries: CE - carotid endarterectomy, LBG - lower extremity bypass graft, MVR - mitral valve replacement, AVR - aortic valve repair, CABG - coronary artery bypass grafting.

4.6 Managerial Implications

To address Questions 2 and 3 from the Introduction, we now turn to examination of the benefits of patient-centric information to patients, payers and providers. To evaluate the impacts on patients, we compare the sets of best hospitals and potential outcomes under population-average and patient-centric information. To illustrate the potential benefit to payers, we use the Hospital Acquired Condition Reduction Program as an example of how patient-centric information enables payers to better align payments with hospital performance. To illustrate the benefits to providers, we discuss how patient-centric information can help hospitals better align their strategic focus with their strengthes and focus their process improvement efforts where they will have the greatest impact.

4.6.1 Implications for Patients

Existing hospital rating systems, such as those of US News and the LeapFrog Group, and quality report cards, such as the New York Cardiac Surgery Quality Report Cards, compare hospitals using O/E ratios of observed to expected metrics (e.g., mortality rate). The expected rates are population averages estimated from a multivariate logit/probit model that includes patient demographics and comorbdities to control for patient severity of illness and hospital dummies to capture the fixed effects of individual hospitals. US News aggregates ratings into broad categories such as heart surgery and cancer, rather than reporting them for individual procedures such as mitral valve or aortic valve surgeries. As a result, it captures only the average effect of a hospital for all discharged patients. The LeapFrog Group and NY quality report cards report ratings for individual procedures such as CABG, mitral valve, aortic valve surgeries, so they capture the average effect of a hospital for a procedure. But they still make use of population-average O/E ratios that do not capture the heterogeneity of outcome differences across groups of patients undergoing the same procedure.

Because population-average based rankings assume away heterogeneity in provider performance across patient groups, they suggest that the same hospitals (or surgeons or physicians) are best for all patients. This leads to two problems. First, as we discussed in the previous section, some hospitals that are high performers on average have average or below average outcomes for some patient groups. So, O/E ratios will guide some patients to suboptimal choices of providers. Second, because they suggest a "one size fits all" picture of hospital quality, population-average based rankings encourage patients to concentrate unnecessarily in a small subset of hospitals. The resulting capacity overloads will lead to longer patient wait times that could negatively impact patient outcomes.

4.6.1.1 Comparison of Best Hospitals

To illustrate the difference between patient-centric and population-average information in terms of their ability to guide patients to the best hospitals, we use each type of information to identify the best hospital(s) (i.e., those that achieve the minimum complication score) for each patient group. The difference between the average complication score under patient-centric and population-average information is a measure of the expected incremental value of patient-centric information to a randomly selected patient who chooses the best hospital for him/her based on the available information.

Because the dependent variable (complication score) is left truncated at zero, we use a tobit model instead of a logit/probit model to identify the best hospital under population-average information. For all models, we have robust standard errors clustered by hospital to allow for differences in the variance/standard errors due to arbitrary intra-group correlation (Jaeker and Tucker 2016, KC and Terwiesch 2011). The hospital with the smallest O/E ratio is designated as the best hospital for all patient groups. To rank hospitals using patient-centric information, we use the instrumental variable tree approach discussed earlier. As we noted earlier, this method can identify different hospitals as best for different patient groups. Furthermore, if the outcome differences between hospitals are not significant, the tree may not differentiate between them. As a result, multiple hospitals may be identified as best for a given patient group.

Applying these methods to data for NY patients discharged in 2012 after one of the six cardiovascular surgeries listed earlier generates the results in Table 4.5. These identify the set of best hospitals (Column 1) and the number of patients for whom each hospital is best under population-average (Columns 2 and 6) and patient-centric information (Columns 3 and 7). The difference in hospital rankings, and the patient complication scores they produce (Columns 4 and 8 for absolute change, and Columns 5 and 9 for relative change), that occur when we switch from population-average information to patient-centric information, characterize the value of patient-centric information to an individual patient who seeks out the best hospital for him/her using the available information. In addition to guiding patients to hospitals that will reduce their expected complication score, patient-centric information guides patients to a wider range of hospitals, which will be more feasible from a capacity standpoint to provide patients with the best available treatment.

4.6.1.2 Comparison of Patient Outcomes

There are two main insights from Table 4.5. The first is that the hospital that is best on average across the entire population is not best for most patients. Patient-centric information reveals that different hospitals are best for different patients. For most of the surgical procedures, the top-ranked hospital under population-average information is the top hospital only for a minority of patients.

For AVR, the top-ranked hospital under population-average information is only best for 512 out of 3,979 patients. For MVR, it is optimal for 571 out of 1,026 patients. For AAA, it is optimal for 63 out of 184 patients. For LBG, the top-ranked hospital under population-average information is only the best for 50 out of 2,324 patients. And for CABG, it is optimal for 1,539 out of 7,826 patients. For CE, it is optimal for 613 out of 2,671 patients.

The second insight from Table 4.5 is that choosing the best hospital on the basis of patient-centric, rather than population-average, information results in a substantial reduction in average complication score. This reduction ranges from 0.04 to 0.37 (or 24.3% to 94.1%) across the six cardiac specialties. The average reduction across all patients is 0.15 (or 66.7%).

To get a better sense of which patient groups benefit most from patient-centric information, we group patients by procedure type, age group and major comorbidities (as what we did for the earlier heat maps). The average reduction of complication score for each patient group is summarized in Figure 4.2. Generally speaking, LBG patients benefit the most, followed by AVR and CABG patients. CE and MVR patients with diabetes, chronic heart failure or no comorbidities benefit the least.

Figure 4.3 displays the distribution of percentage reduction in complication score. From this histogram, we see that around 80% of patients achieve a positive reduction in their complication score under patient-centric information. A large majority of them achieve 70-90% reduction in their complication score. Hence, the answer to Question 2 in the Introduction is that patient-centric data is highly valuable to a strong majority of cardiovascular surgery patients.

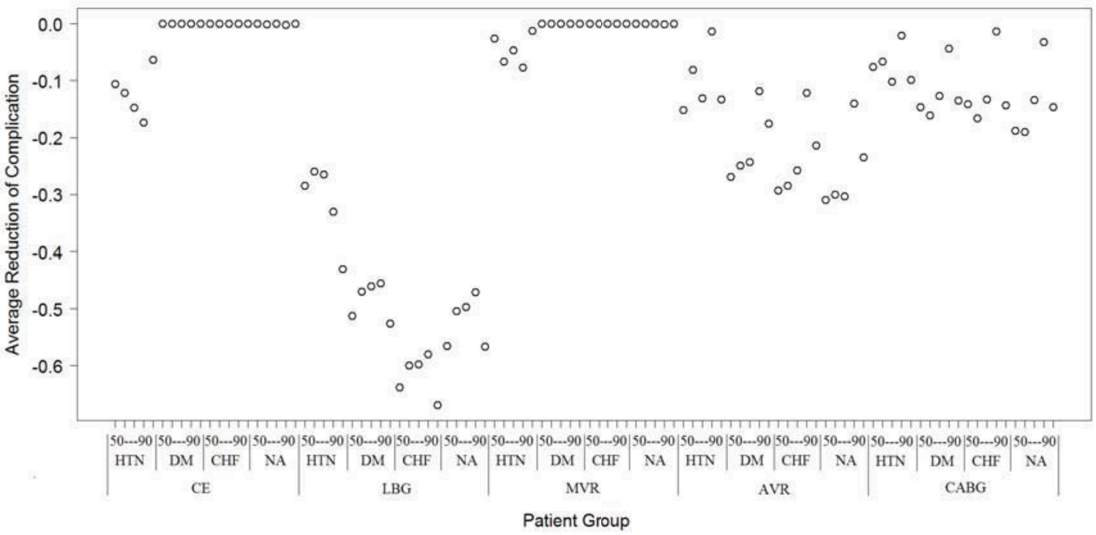
4.6.2 Implications for Hospitals and Payers

Payers are increasingly seeking ways to tie hospital reimbursement to performance. For example, the Hospital Acquired Condition Reduction Program (HACRP) was established in 2013 as a response to increasing costs of complications. This program penalizes low-performing hospitals with regard to the Patient Safety Indicator (PSI)

Table 4.5: Complication Reduction from Using Patient-Centric Instead of Population-Average Information in Hospital Selection

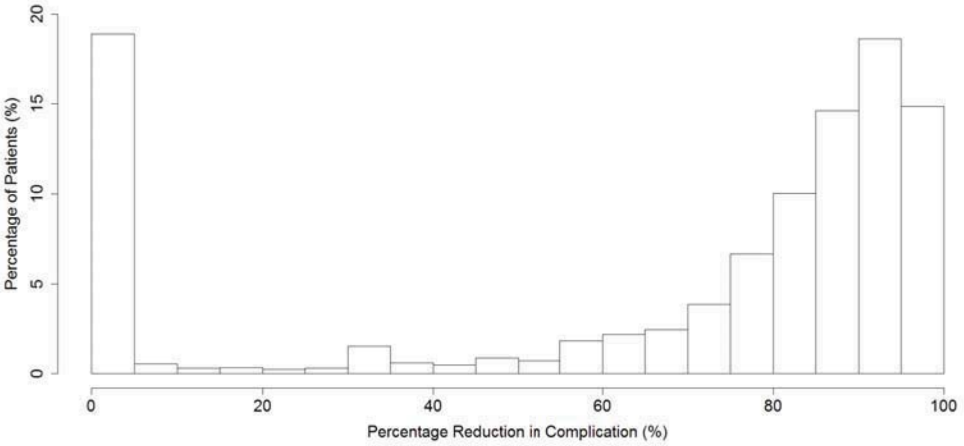
Hospital Index	Number of Patients		Change in Complication Score (absolute)	Change in Complication Score (relative)	Number of Patients		Change in Complication Score (absolute)	Change in Complication Score (relative)
	population-average	patient-centric			population-average	patient-centric		
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	AVR				MVR			
1		2,804	-0.19	82.4%	1,026	571	0	0%
3	3,979	512	0	0%		149	-0.06	60.8%
5		71	-0.16	35.4%		40	-0.20	30.6%
17		398	-0.22	75.4%		237	-0.07	53.6%
20		23	-0.03	47%		21	-0.11	70.3%
25		170	-0.12	85.5%		6	-0.08	71.2%
32		1	-0.25	26.6%		2	-0.09	7.8%
Overall			-0.16	70.2%			-0.04	24.3%
	AAA				LBG			
1	184	63	0	0%		1,535	-0.41	96.5%
3		10	-0.05	47.4%		114	-0.32	97.0%
5		2	-0.17	19.3%	2,324	50	0	0%
17		70	-0.10	51.2%		391	-0.34	94.6%
20						8	-0.24	94.4%
25		39	-0.10	59.8%		220	-0.29	96.6%
27						6	-0.29	95.7%
Overall			-0.06	31.5%			-0.37	94.1%
	CABG				CE			
1		2,522	-0.17	84.1%	2,671	613	0	0%
3	7,826	1,539	0	0%		1	-0.03	77.2%
4		344	-0.04	62.3%		1	-0.01	49.5%
5		67	-0.14	36.0%		185	-0.41	77.5%
6		3	-0.16	27.9%				
7		1	-0.14	38.0%				
9						7	-0.02	59.2%
11		8	0.00	8.6%				
17		1,348	-0.16	76.7%		1,240	-0.14	88.3%
20		57	-0.08	79.5%		5	-0.13	92.4%
25		1,936	-0.09	76.4%		616	-0.10	89.3%
27						3	-0.07	83.1%
32		1	-0.23	26.2%				
Overall			-0.11	62.8%			-0.11	67.4%

Figure 4.2: Complication Reduction by Patient Group



Note: Patients are grouped by age group (i.e., 50s to 90s), comorbidity and surgery. Acronyms for comorbidities: HTN - hypertension, DM - diabetes, CHF - chronic heart failure, NA - no comorbidities. Acronyms for surgeries: CE - carotid endarterectomy, LBG - lower extremity bypass graft, MVR - mitral valve replacement, AVR - aortic valve repair, CABG - coronary artery bypass grafting.

Figure 4.3: Complication Reduction under Patient-Centric Information



90 Composite Index Value (Domain 1) and five infection measures (Domain 2).[8] For each measure, CMS uses two years of historical data to calculate risk-adjusted infection rates and then ranks hospitals accordingly. Each hospital is assigned a score between 1 and 10 for each measure based on its relative rank in deciles for that measure. There is only one score for Domain 1. A hospital's Domain 2 score is calculated as the average of the domain's individual measures. The total score is calculated as the weighted average of Domain 1 and Domain 2 scores, where the weights are 15% and 85% for the two domains. In 2015, CMS reduced total payments (i.e., across all patients) by 1% for hospitals that ranked among the worst quartile with regard to hospital acquired infections.

4.6.2.1 Impact of Patient-Centric Information on Hospital Payments

The Hospital Acquired Condition Reduction Program is based on population-average outcome information and so does not recognize heterogenous outcome differences across patient groups. Consequently, applying a uniform penalty to these hospitals does not recognize their acceptable or even high performance for some patient groups. Similarly, hospitals that are not penalized under the HACRP may perform poorly for some patient groups. In addition to misaligning penalties with performance, an incentive system based on population-average information can hide areas of poor performance and discourage hospitals from addressing them. In contrast, patient-centric information allows payers to assess hospital performance by patient group and better align payments with quality to provide shaper incentives for quality improvement.

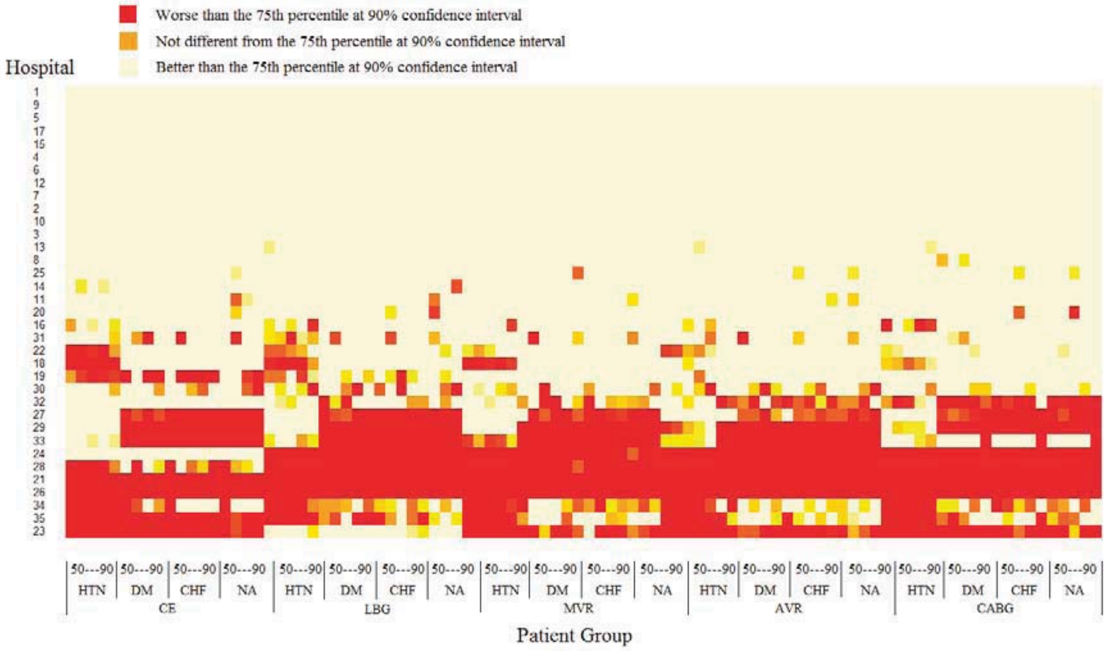
To illustrate a HACRP-type program under patient-centric information, we group patients by procedure type, age group and comorbidities. For each patient group, we use $Y_{ijk} \in \{0, 1\}$ to indicate whether hospital j is among the worst quartile for patient i in group k. We then calculate the overall performance of hospital j for patient group k using $\bar{Y}_{jk} = \frac{1}{N_{jk}} \sum_{i=1}^{N_{jk}} Y_{ijk}$ and display the results in the heat map of Figure 4.4. We see that only Hospitals 21 and 26 are among the worst quartile across all patient groups. Hospitals 23, 27, 28, 29 and 33 are among the worst quartile for a majority of patient groups, but they have areas (e.g., procedure LBG for Hospital 23) that are not among the worst quartile. Likewise, Hospitals 11, 14 and 25 are not among the worst quartile

[8] The PSI measures include rates of pressure ulcer, iatrogenic pneumothorax, central venous catheter-related bloodstream infection, postoperative hip fracture, perioperative pulmonary embolism or deep vein thrombosis, postoperative sesis, postoperative wound dehiscence and accidental puncture or laceration. The five infection measures are rates of central line-associated bloodstream infection, catheter-associated urinary tract infection, colon and hysterectomy surgical site infection, methicillin-resistant staphlococcus aureus bacteremia, and clostrium dfficile infection.

for the majority of patient groups, but they have areas (e.g., old MVR patients with diabetes for Hospital 25) that are among the worst quartile.

Payments would be better aligned with performance if hospitals were penalized for only their low-performing areas. To see how, in Figure 4.5, we compare scenarios in which hospitals are penalized based on population-average and patient-centric information. Under population-average information, there are nine hospitals with average performance among the worst quartile, each of which would be penalized by 1% on all payments. The other hospitals are not penalized at all. In contrast, under patient-centric information, only ten hospitals are not penalized at all. The rest are penalized on some portion of their payments. Hence, more hospitals would have a financial incentive to improve under patient-centric than under population-average information.

Figure 4.4: Comparison of Hospitals' Performance by Patient Group

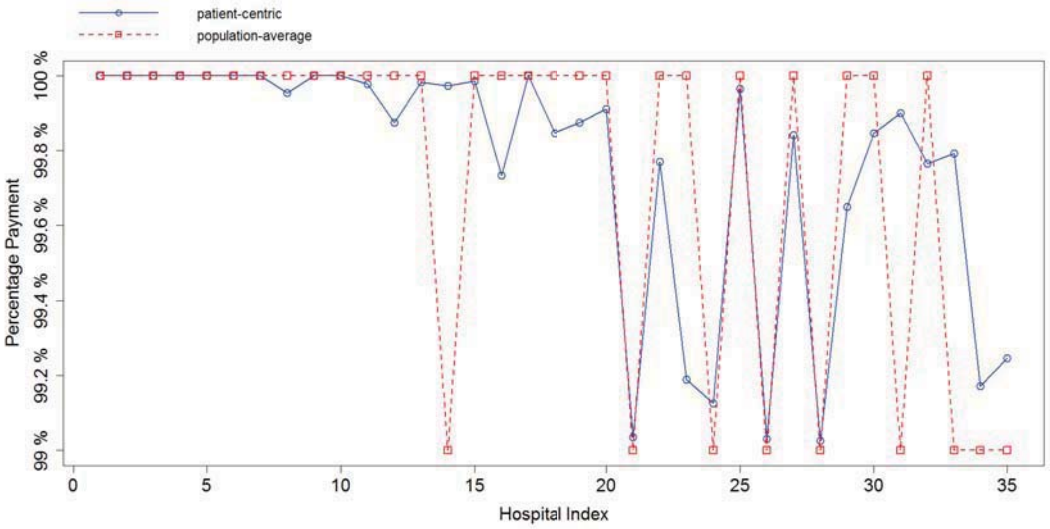


Note: Patients are grouped by age group (i.e., 50s to 90s), comorbidity and surgery. Acronyms for comorbidities: HTN - hypertension, DM - diabetes, CHF - chronic heart failure, NA - no comorbidities. Acronyms for surgeries: CE - carotid endarterectomy, LBG - lower extremity bypass graft, MVR - mitral valve replacement, AVR - aortic valve repair, CABG - coronary artery bypass grafting.

4.6.2.2 Impact on Hospital Strategy and Improvement Efforts

Payments based on patient-centric information provide more focused incentives for hospitals to improve quality, because they reward hospitals for incremental improvements. For example, consider a hospital that discharges 1,000 patients a year, of which 100 are CABG patients. The infection rate across all patients is 1%, but is 5% for CABG pa-

Figure 4.5: Percentage Payment under Patient-Centric and Population-Average Measure



tients. If, under the current HACRP, the hospital is not penalized, it has no economic incentive to improve. Even if it is penalized, it may be the case that reducing infections among CABG patients will not have a large enough effect on the overall infection rate to eliminate the penalty. However, if HACRP penalties were based on patient-centric information, and therefore individually penalized payments for CABG patients, then the hospital would have economic incentives to reduce the CABG patient infection rate, regardless of whether payments for other types of patients were being penalized or not.

Beyond its use in targeted incentives, transparent patient-centric outcome information can help hospitals learn from one another. For example, the heat map in Figure 4.4 shows that Hospital 27 has very low complication scores for hypertension patients, despite having average performance for other patients. This may indicate that Hospital 27 has made some kind of innovation that enables them to better protect these patients. Hence, patient-centric information in Figure 4.4 can help hospitals spot best practices that might be shared to elevate performance across the industry.[9]

Finally, in addition to supporting incentives for hospitals to improve outcomes for specific patient groups, patient-centric information may also incent hospitals to focus

[9] Competition may hinder sharing of best practices across hospitals. But there are platforms for such sharing. For example, the Quality Collaborative of the Michigan Society of Thoracic Surgeons http://mstcvs.org/qc.html has been set up precisely to encourage the open heart programs in the state of Michigan to share data and practices.

on the patients they are able to treat most successfully. For example, suppose a hospital has exceptionally good outcomes (e.g., low complication scores), relative to the state average, for elderly patients, but poor outcomes for younger patients. The penalties from an HACRP-type program would make the younger patients less economically attractive to the hospital. And, if patient-centric information were transparently available to patients, demand from younger patients would presumably be weaker as well. Both factors would encourage the hospital to focus on elderly patients, in its process design and marketing efforts. Other hospitals might be incented to focus on particular medical procedures or patient groups (e.g., patients with hypertension, diabetes or cancer). Over time, this would encourage a network of providers that leverage their individual strengths to produce better patient outcomes. These observations indicate the answers to Question 3 in the Introduction is that patient-centric information enhances the power of pay-for-performance systems and sharpens the ability of providers to make quality improvements.

4.7 Conclusion

In recent years, there have been many wide-ranging efforts to improve the delivery of health care in the United States. Perhaps the most straightforward of these has been the push for better and more transparent outcome information to help patients find the best available care for them. Unfortunately, as we have shown, the standard approach of computing risk-adjusted outcomes produces population averages that do not accurately represent the likely outcomes for all patients. In this paper, we have shown that the relative performance of hospitals is heterogeneous across patient groups. Consequently, patient-centric rankings of hospitals are significantly different than rankings based on population-average information.

In this study, we have addressed the challenges of generating patient-centric outcome information and hospital ranking. Using six cardiovascular surgeries as the clinical setting, we studied the outcomes of thirty-five hospitals in New York state. We extended the instrumental variable tree approach for multiple hospitals to recursively partition patients into groups that exhibit significant outcome differences between hospitals. We quantified the outcome differences for groups of patients using the instrumental variable method and derived patient-centric estimates of outcome differences between hospitals for individual patients. Our analysis shows that outcome differences between hospitals are heterogeneous not only across procedure types, but also along other dimensions such as patient age and comorbidities.

We compared the best hospitals based on population-average and patient-centric information. We found that, for the majority of patients (around 80%), the best hospitals are different than those indicated as best by a population-average rating. Furthermore, we found that patient-centric information results in a larger set of best hospitals, which suggests more opportunities for distributing patient workload across hospitals to reduce patient waiting time. Most importantly, we compared the potential outcomes when patients are treated at the best hospitals based on the two types of information, and estimated that the complication score could be reduced by 66.7% by using patient-centric information instead of population-average information.

In addition to the manifest benefits to patients, patient-centric information offers potential benefits to hospitals and payers as well. Using the Hospital Acquired Infection Reduction Program as an example, we showed that patient-centric information allows the CMS to better align payments (and penalties) with patient outcomes. This in turn provides sharper incentives for hospitals to improve quality. Finally, the more detailed patient-centric information can help hospitals to understand their strengths and weaknesses, as well as those of their peers. This can help them better align their strategies with their strengths, and also to learn from one another.

APPENDIX A

Appendix to Chapter 2

A.1 Hausman Test Results

Table A.1: Hausman Test Results

	Fixed Effects (F)	Random Effects (R)	Difference $(F-R)$	Std. Err. $\sqrt{diag(V_F - V_R)}$
Surgical Volumes				
hosp_vol	0.56	0.21	0.35	0.52
surg_vol	−3.29	0.13	−3.42	7.80
Patient Characteristics				
age	−0.02	−0.02	0.00	0.00
female	−0.34	−0.31	−0.03	0.02
black	−0.09	−0.17	0.08	0.05
hispanic	−0.23	−0.32	0.09	0.06
asian	−0.26	−0.29	0.04	0.10
others	0.17	0.13	0.04	0.04
Comorbidities				
atrial fibrillation	−0.11	−0.12	0.01	0.02
alcohol abuse	0.38	0.28	0.09	0.09
deficiency anemias	0.07	0.09	−0.03	0.03
rheumatoid arthritis	−0.31	−0.27	−0.04	0.05
blood loss	0.16	0.08	0.08	0.13
heart failure	0.19	−0.09	0.28	0.21
lung disease	−0.25	−0.24	−0.01	0.02
coagulopathy	−0.11	−0.13	0.02	0.02
depression	0.08	0.07	0.01	0.04
diabetes	−0.10	−0.12	0.02	0.02
drug abuse	−0.17	−0.11	−0.06	0.08
hypertension	0.05	0.05	0.00	0.02
hypothyroidism	0.09	0.07	0.02	0.03
liver disease	−0.31	−0.15	−0.16	0.12
lymphoma	−0.22	−0.22	0.00	0.10
electrolyte disorders	0.07	0.04	0.03	0.02
metastatic cancer	−0.61	−0.53	−0.08	0.12
neurological disorders	0.11	0.10	0.02	0.05
obesity	−0.15	−0.20	0.05	0.03
paralysis	−0.12	−0.11	−0.01	0.07
vascular disorders	0.03	0.00	0.03	0.04
psychoses	−0.26	−0.31	0.05	0.10
pulmonary disorders	−0.47	−0.25	−0.22	0.25
renal failure	−0.33	−0.29	−0.04	0.03
solid tumor	−0.22	−0.19	−0.04	0.08
valvular disease	−0.52	−0.47	−0.05	0.30
weight loss	−0.01	−0.05	0.04	0.06
constant	0.01	0.79	−0.79	1.14

Note: H_o = difference in coefficients not systematic. F = consistent under H_o and H_a; obtained from probit. R = inconsistent under H_a, efficient under H_o; obtained from multilevel probit.
Test results: $\chi^2(36) = (F-R)'[(V_F - V_R)^{-1}](F-R) = 33.78$, $Prob > \chi^2 = 0.5747$.

A.2 Population-Average Rates of Complication and Readmission

Figure A.1: Complication Rate by Surgeon for A Patient with Average Characteristics

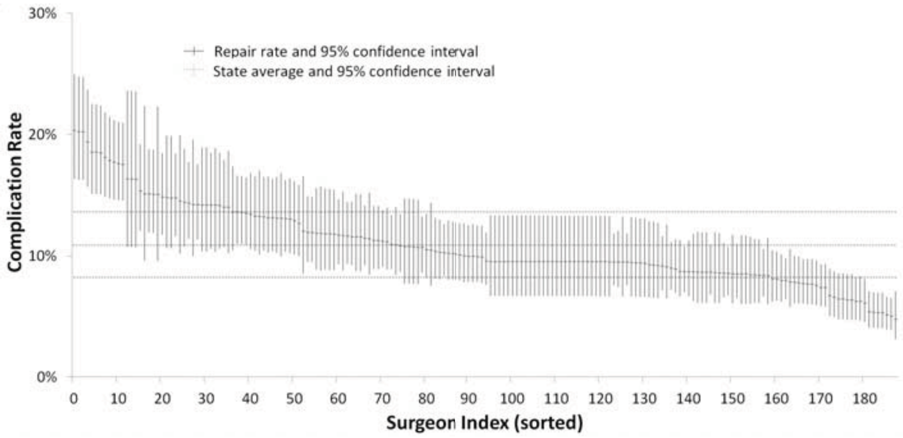
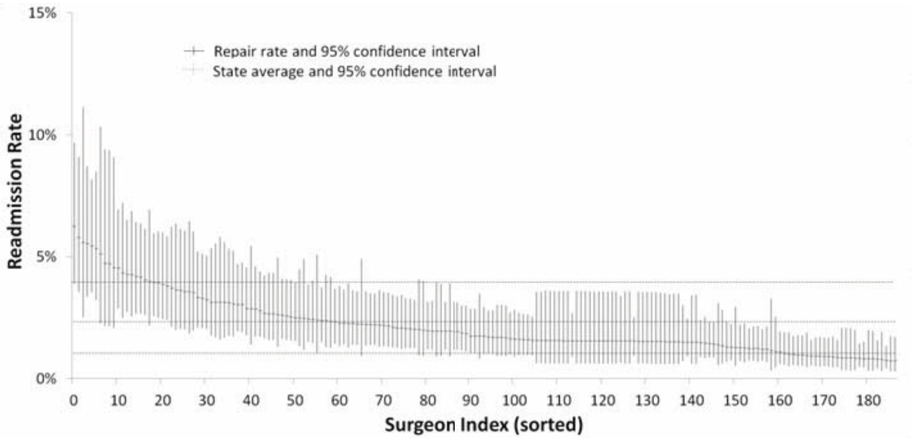


Figure A.2: Readmission Rate by Surgeon for A Patient with Average Characteristics



A.3 Patient-Specific Rates of Complication and Readmission

Figure A.3: Complication Rate by Surgeon for Patients of Different Levels of Acuity

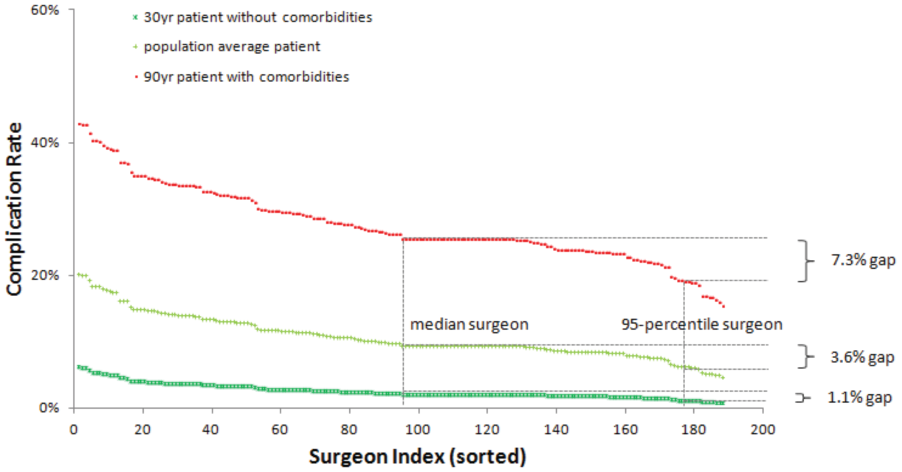
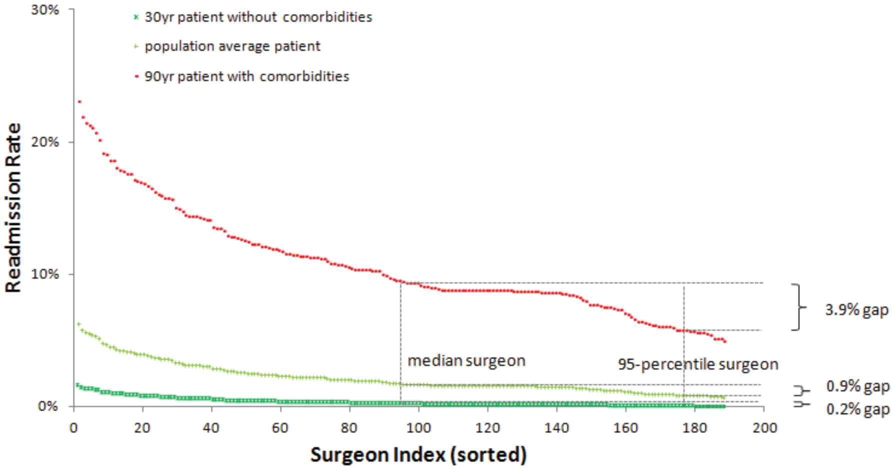


Figure A.4: Readmission Rate by Surgeon for Patients of Different Levels of Acuity



A.4 Estimation Results of Each Type of Complication

Table A.2: Estimation Results of Each Type of Complication

	Stroke		Wound Infection		Renal Failure		Reoperation		Ventilation	
	Coeff.	Std. Err.	Coeff.	Std. Err.	Coeff.	Std. Err.	Coeff.	Std. Err.	Coeff.	Std. Err.
Surgical Volumes										
hosp_vol	−0.21**	0.10	0.12	0.15	0.04	0.06	−0.24**	0.10	0.31***	0.11
surg_vol	0.16***	0.04	0.06	0.06	−0.05**	0.03	0.02	0.03	−0.08***	0.02
Patient Demographics										
age	0.02*	0.01	0.00	0.01	0.02***	0.00	0.02***	0.00	0.02***	0.01
female	0.17	0.22	0.20*	0.10	−0.12*	0.06	0.24**	0.11	0.18*	0.10
black	0.27	0.52	−0.43	0.47	0.35**	0.17	−0.07	0.25	0.16	0.20
hispanic	0.79**	0.34	−0.20	0.17	0.02	0.15	−0.06	0.24	0.11	0.21
asian	0.57	0.42	−5.07***	0.38	−0.19	0.18	0.05	0.39	−0.18	0.45
others	0.47**	0.24	−0.05	0.26	−0.16	0.13	−0.36	0.23	−0.20	0.26
Comorbidities										
atrial fibrillation	−0.02	0.17	−0.14	0.18	−0.07	0.08	0.04	0.16	−0.19	0.12
alcohol abuse	−5.23***	0.83	−4.35***	0.40	−0.09	0.23	−5.16***	0.33	−0.03	0.58
deficiency anemias	−0.09	0.33	−0.45**	0.22	−0.18	0.13	0.12	0.09	−0.46**	0.20
rheumatoid arthritis	−4.46	.	0.10	0.44	−0.23	0.24	−0.38	0.49	−4.91***	0.29
blood loss	−4.07	.	−4.44***	1.05	−5.10***	0.53	0.20	0.49	−4.08***	0.19
heart failure	0.44	0.34	0.32	0.82	0.29	0.67	−6.15***	0.51	3.03***	0.80
lung disease	0.03	0.23	0.12	0.22	0.02	0.10	0.01	0.12	−0.01	0.12
coagulopathy	−0.28	0.26	−0.12	0.15	0.32***	0.04	0.28**	0.12	0.28***	0.10
depression	−0.60	0.40	−0.05	0.42	−0.19	0.18	−0.32	0.33	−0.35	0.26
diabetes	−0.20	0.24	0.36**	0.14	0.07	0.12	−0.11	0.12	0.07	0.16
drug abuse	−3.49	.	−4.46***	0.66	−0.17	0.63	−4.71***	0.27	0.71*	0.39
hypertension	−0.54**	0.24	−0.21	0.15	−0.28***	0.09	−0.26***	0.08	−0.37***	0.13
hypothyroidism	0.05	0.25	−0.46	0.33	−0.08	0.13	0.10	0.14	−0.13	0.32
liver disease	−4.85***	0.76	0.41	0.67	0.66***	0.25	0.75**	0.32	−1.15**	0.48
lymphoma	−4.72	.	−4.81***	0.36	0.90**	0.39	0.19	0.49	−4.30***	0.23
electrolyte disorders	−0.15	0.25	−0.07	0.18	0.42***	0.12	0.14	0.15	0.14	0.10
metastatic cancer	−3.67	.	−4.36***	1.02	−4.58***	0.21	−4.90***	0.32	−4.13***	0.21
neurological disorders	0.71**	0.36	−4.79***	0.33	−0.44	0.33	0.05	0.33	0.25	0.20
obesity	0.37	0.36	0.03	0.27	0.15	0.18	−0.25*	0.15	−0.03	0.23
paralysis	2.09***	0.32	0.62	0.42	−0.18	0.33	0.61*	0.32	0.74***	0.25
vascular disorders	−0.55**	0.22	0.06	0.26	0.15	0.20	−0.39	0.29	0.42**	0.18
psychoses	−4.24	.	0.75	0.47	−0.31	0.48	−4.84***	0.14	0.03	0.40
pulmonary disorders	0.73	0.59	−5.43***	0.89	0.17	0.76	−6.16***	0.55	6.94***	0.26
renal failure	0.26	0.23	0.62***	0.18	1.11***	0.11	−0.09	0.15	0.35*	0.21
solid tumor	−3.98	.	0.73***	0.28	−0.06	0.41	−4.80***	0.37	−4.54***	0.12
valvular disease	0.14	0.42	0.83	0.94	0.50	0.46	1.37***	0.47	1.25***	0.44
weight loss	0.76***	0.29	0.62**	0.28	0.56***	0.12	0.69***	0.16	1.42***	0.18
Others										
repair	−0.20	0.21	−0.46**	0.24	−0.18*	0.09	−0.11	0.13	−0.24***	0.09
alpha	0.00	0.00	0.00	0.08	0.00	0.00	0.12	0.06	0.11	0.07
beta	0.00	0.00	0.02	0.05	0.00	0.00	0.02	0.05	0.00	0.00
constant	−3.32***	0.84	−3.21***	0.88	−3.08***	0.40	−2.22***	0.40	−4.45***	0.65
log likelihood	−74.37		−115.87		−594.03		−359.75		−244.36	

Note: *** p < 0.01, ** p <0.05, * p < 0.1. Robust standard errors are clustered by surgeon.

A.5 Model Elements and Sources for the Estimation of QALE

Since there is no single source or paper that provides the value of various parameters in our model, we draw from several sources in the medical literature to estimate long-term risks associated with mitral valve repair and replacement and quality of life associated with different risks. Below we discuss our estimates and sources of each model element for patients of different ages and different comorbidities.

Stroke: We estimate the risk of stroke based on Russo et al. (2008), who studied 1,344 patients that underwent mitral surgery at the Mayo Clinic from Jan 1980 to Dec 1995, and reported that (1) annual rate of stroke is 1.15% for mitral valve repair, 1.65% and 2.7% for biological and mechanical valve replacement, respectively, (2) risk ratio of age is 1.1 per 5 years, and (3) risk ratio of atrial fibrillation is 1.4 for both mitral vale repair and biological valve replacement. We estimate quality of life after stroke based on Shah and Gage (2011), who developed a decision-analysis model to compare the cost and quality-adjusted survival of various antithrombotic therapies on the basis of the results from Randomized Evaluation of Long Term Anticoagulation Therapy and other trials, and reported that quality of life after moderate to severe stroke is 0.39.

Bleeding: We estimate the risk of bleeding based on Russo et al. (2008), who reported that (1) annual rate of bleeding is 0.7% for mitral valve repair, 1.4% and 2.43% for biological and mechanical valve replacement, respectively, (2) risk ratio of age is 1.14 per 5 years, and (3) risk ratio of atrial fibrillation is 1.52 for both mitral vale repair and biological valve replacement. We estimate quality of life after stroke based on Shah and Gage (2011), who reported that quality of life after bleeding is 0.8.

Structural Valve Deterioration (SVD): We estimate the risk of structural valve deterioration based Bourguignon et al. (2014), who studied 450 patients that underwent biological valve replacement from 1984 to 2011, and reported that annual rate of structural valve deterioration is 2.75%, 1.5%, 0.6%, 0.4% for patients at their 50s, 60s, 70s and 80s, respectively. Because structural valve deterioration usually requires reoperation (Bourguignon et al. 2014), we assume that quality of life after structural valve deterioration is similar to that after reoperation.

Long-term Survival: We estimate long-term survival of patients with mitral valve repair based on the US Social Security database, assuming that mitral valve repair restores patients' normal life expectancy (Ray et al. 2006). Long-term survival associated with mitral valve replacement is estimated based on Daneshmand et al. (2010), who studied 2,064 patients that underwent isolated primary mitral operations, and found that (1) annual mortality rates associated with biological and mechanical valve replacement are 1.8 and 1.3 times that associated with mitral valve repair, and (2) risk ratio is 1.4 for diabetes and 1.3 for lung disease. In a similar study, Daneshmand et al. (2009) found that risk ratio is 2.68 for renal disease and 1.37 for hypertension. Risk ratio of other comorbidities is estimated to be 1.6 for heart failure (Gelsomino et al. 2011) and 2.3 for atrial fibrillation (Ruel et al. 2004).

Lastly, quality of life is estimated to be 0.6 for readmission (Cox et al. 2007), 0.45 for reoperation (Regier et al. 2006), 0.7 for ventilation (Windisch et al. 2003) and 0.85 for wound infection (Jidéus et al. 2009). Quality of life for patients with comorbidities is estimated to be 0.751 for diabetes, 0.636 for heart failure, 0.714 for lung disease, 0.651 for renal failure, 0.789 for hypertension and 0.774 for atrial fibrillation (Sullivan and Ghushchyan 2006).

A.6 Numerical Analysis Results with Heterogeneous Weights on Travel Distance

Table A.3: Comparison of the Effectiveness of Patient-Specific Information and Capacity Increase

Weight on Distance (1)	Weight on Waiting Time (2)	Expected Number of Repairs (3)	Average Distance (miles) (4)	Average Waiting (months) (5)	Convenience Adjusted QALE*(days) (6)	Expected Number of Repairs (7)	Average Distance (miles) (8)	Average Waiting (months) (9)	Convenience Adjusted QALE*(days) (10)
		Patient-Specific (current capacity)				Population-Average (current capacity)			
Low	Low	2,174	26	5.7	258	2,104	26	5.8	243
	Medium	2,163	28	1.1	257	2,090	28	1.1	240
	High	2,153	30	0.5	256	2,099	30	0.5	243
Medium	Low	2,163	22	5.6	246	2,099	22	5.7	232
	Medium	2,152	23	1.1	244	2,095	24	1.1	231
	High	2,147	24	0.5	245	2,088	24	0.5	232
High	Low	2,115	17	5.2	171	2,064	17	5.3	158
	Medium	2,101	17	1.0	168	2,054	17	1.1	155
	High	2,087	17	0.5	168	2,044	17	0.5	156
		Population-Average (10% capacity increase)				Population-Average (20% capacity increase)			
Low	Low	2,122	26	5.8	248	2,131	26	5.4	255
	Medium	2,115	28	1.1	249	2,133	29	1.1	254
	High	2,116	30	0.5	249	2,135	30	0.5	256
Medium	Low	2,116	22	5.4	240	2,129	23	5.3	245
	Medium	2,111	24	1.1	238	2,131	24	1.1	244
	High	2,108	25	0.5	238	2,127	25	0.5	245
High	Low	2,082	17	4.9	167	2,102	17	4.6	175
	Medium	2,074	17	1.0	164	2,095	17	0.9	171
	High	2,063	17	0.5	163	2,088	18	0.5	170
Actual		1,557	19						

Note: This table compares scenarios when patients choose surgeons based on patient-specific information (with current capacity) and population-average information (with 0–20% capacity increases). We consider Low, Medium and High weights patients place on travelling and waiting. Equivalent quality-adjusted life days per mile for Low, Medium and High weights on travelling are 0.5, 1, 5. Equivalent quality-adjusted life days per month for Low, Medium and High weights on waiting are 10, 50, 100. *For the ease of comparison, a fixed amount of one quality-adjusted life year has been subtracted from Convenience Adjusted QALE for both patient-specific and population-average cases.

Table A.4: Comparison of the Values from Using Patient-Specific and Population-Average Information

	Diff. in Num. of Repairs			**Diff. in Total QALE (year)**			**Diff. in Average Travel Dist.(mile)**			**Diff. in Average Wait Time(month)**		
Weight on	Weight on Waiting			Weight on Waiting			Weight on Waiting			Weight on Waiting		
Distance	Low	Medium	High	Low	Medium	High	Low	Medium	High	Low	Medium	High
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
Low	71	74	55	108	120	81	0	0	0	-0.1	0.0	0.0
Medium	64	57	59	93	84	89	0	0	0	-0.2	0.0	0.0
High	51	47	43	70	65	59	0	0	0	-0.2	-0.1	0.0

Note: This table summarizes the changes in total number of repairs, quality-adjusted life years, average travel distance and waiting time per patient when information is switched from population-average to patient-specific. We consider Low, Medium and High weights patients place on travelling and waiting. Equivalent quality-adjusted life days per mile for Low, Medium and High weights on travelling are 0.5, 1, 5. Equivalent quality-adjusted life days per month for Low, Medium and High weights on waiting are 10, 50, 100.

APPENDIX B

Appendix to Chapter 3

B.1 Estimation of Mean Squared Errors

Let $\beta(X_i)$ denote the conditional average treatment effect for subject i. We have $E_{S^{es}}[\hat{\beta}^{es}(X_i)|S^{es}] = E_{X_i}[\beta^{te}(X_i)] = \beta(X_i)$. The expected MSE is the expectation of $MSE(S^{te}, S^{es})$ over test and estimation samples:

$$\begin{aligned}
EMSE &= E_{S^{te},S^{es}} MSE(S^{te}, S^{es}) \\
&= E_{S^{te},S^{es}}[(\beta^{te}(X_i) - \hat{\beta}^{es}(X_i))^2] \\
&= E_{S^{te},S^{es}}[(\beta^{te}(X_i) - \beta(X_i))^2 + \hat{\beta}^{es}(X_i)^2 - \beta(X_i)^2 + 2\beta^{te}(X_i)(\beta(X_i) - \hat{\beta}^{es}(X_i))] \\
&= E_{S^{te},S^{es}}[(\beta^{te}(X_i) - \beta(X_i))^2 + \hat{\beta}^{es}(X_i)^2 - \beta(X_i)^2 + 2\beta(X_i)(\beta(X_i) - \hat{\beta}^{es}(X_i))] \\
&= E_{S^{te},S^{es}}[(\beta^{te}(X_i) - \beta(X_i))^2 + (\hat{\beta}^{es}(X_i) - \beta(X_i))^2] \\
&= E_{S^{te}}[(\beta^{te}(X_i)^2 - 2\beta^{te}(X_i)\beta(X_i) + \beta(X_i)^2] + E_{X_i,S^{es}}[(\hat{\beta}^{es}(X_i) - \beta(X_i))^2] \\
&= E_{S^{te}}[\beta^{te}(X_i)^2 - \beta(X_i)^2] + E_{X_i,S^{es}}[Var(\hat{\beta}^{es}(X_i))].
\end{aligned}$$

Because $E_{S^{te}}[\beta^{te}(X_i)^2]$ does not depend on the estimator, minimizing above EMSE is equivalent to minimizing

$$EMSE(S^{te}, S^{es}) = -E_{X_i}[\beta(X_i)^2] + E_{X_i,S^{es}}[Var(\hat{\beta}^{es}(X_i))].$$

B.2 Proofs of Lemma, Theorem and Corollary

B.2.1 Proof of Lemma 1

The first part of Lemma 1 follows directly from Slutsky theorem for probability limits, which state that, for a continuous function $g(X_n)$, $plim(g(X_n)) = g(plim(X_n))$. To prove the second part, we use the Dominated Convergence theorem, which states that, if $X_N \xrightarrow{p} X$ and there is a random variable Z with $E[Z] < \infty$ such that $|X_N| < Z$ for all N, then $E[\lim_{N\to\infty} X_N] = \lim_{N\to\infty} E[X_N]$. Therefore, we have

$$\begin{aligned} \lim_{N\to\infty} Cov(\beta_i T_i, Y_i) &= \lim_{N\to\infty}(E[\beta_i T_i Y_i] - E[\beta_i T_{1i}]E[Y_i]) \\ &= \lim_{N\to\infty} E[\beta_i T_i Y_i] - \lim_{N\to\infty} E[\beta_i T_i]E[Y_i] \\ &= E[\lim_{N\to\infty} \beta_i T_i Y_i] - E[\lim_{N\to\infty} \beta_i T_i]E[Y_i] \\ &= E[\beta_{l_j} T_i Y_i] - E[\beta_{l_j} T_i]E[Y_i] \\ &= Cov(\beta_{l_j} T_i, Y_i). \end{aligned}$$

The third equality follows from the observation that $\beta_i T_i Y_i \xrightarrow{p} \beta_{l_j} T_i Y_i$ and $\beta_i T_i \xrightarrow{p} \beta_{l_j} T_i$ (by the Slutsky theorem). Because β_i, T_i and Y_i are all bounded in practice, there exist random variables Z_1 and Z_2 such that $E|Z_1| < \infty$, $E|Z_2| < \infty$, $|\beta_i T_i Y_i| < Z_1$ and $|\beta_i T_i| < Z_2$.

B.2.2 Proof of Lemma 2

We break Lemma 2 into two parts and prove them separately: (a) the moment generating function of V_i is bounded by $e^{\frac{t^2(b-a)^2}{8}}$ (i.e., $\psi_i(t) \leq e^{\frac{t^2(b-a)^2}{8}}$), and (b) for any $w > 0$, we have $Pr(\overline{V} \geq w) \leq N^{-7w^2 k_N/[8(b-a)^2]}$, where a and b are lower and upper bounds of V_i.

B.2.2.1 Proof of Part A

Because subjects in the same group are independent of each other, we assume the variable V_i is independent. By Assumption 3, V_i is bounded by a and b, i.e., $a \leq V_i \leq b$, so we can write V_i as a convex combination of a and b, $V_i = \lambda b + (1-\lambda)a$, where $\lambda = \frac{V_i - a}{b-a}$. Because the function $f(x) = e^{tx}$ is convex, we have

$$e^{tV_i} \leq \lambda e^{tb} + (1-\lambda)e^{ta} = \frac{V_i - a}{b-a}e^{tb} + \frac{b - V_i}{b-a}e^{ta}.$$

Taking expectation of both sides, we have

$$E(e^{tV_i}) \leq \frac{\overline{V}-a}{b-a}e^{tb} + \frac{b-\overline{V}}{b-a}e^{ta}.$$

Let $e^{g(u)} = \frac{\overline{V}-a}{b-a}e^{tb} + \frac{b-\overline{V}}{b-a}e^{ta}$, $\gamma = \frac{\overline{V}-a}{b-a}$, and $u = t(b-a)$. We have

$$\begin{aligned} g(u) &= log(\tfrac{\overline{V}-a}{b-a}e^{tb} + \tfrac{b-\overline{V}}{b-a}e^{ta}) \\ &= log(e^{ta}(\tfrac{\overline{V}-a}{b-a}e^{tb-ta} + \tfrac{b-\overline{V}}{b-a})) \\ &= ta + log(\gamma e^u + (1-\gamma)) \\ &= t\overline{V} - \gamma u + log(\gamma e^u + (1-\gamma)). \end{aligned}$$

Note that $g(0) = t\overline{V}$, $g'(0) = 0$ and $g''(u) \leq \frac{1}{4}$ for all $u > 0$. By Taylor's theorem, there is a $\varepsilon \in (0, u)$ such that

$$g(u) = g(0) + ug'(0) + \frac{u^2}{2}g''(\varepsilon) \leq t\overline{V} + \frac{u^2}{8} = t\overline{V} + \frac{t^2(b-a)^2}{8}.$$

Therefore, we have $E(e^{tV_i}) \leq e^{t\overline{V}+\frac{t^2(b-a)^2}{8}}$. That is, the moment-generating function of V_i is bounded by $e^{t\overline{V}+\frac{t^2(b-a)^2}{8}}$. From the exogeneity condition, we have $\overline{V} = E[Z_i\epsilon_i] = E_{Z_i}[Z_iE(\epsilon_i|Z_i)] = 0$. Therefore, we have $\psi_i(t) \leq e^{\frac{t^2(b-a)^2}{8}}$.

B.2.2.2 Proof of Part B

Let $K = (b-a)^2$. For any $w > 0$, from Markov's Inequality and Part A, we have

$$\begin{aligned} Pr(\overline{V} \geq w) &= Pr(e^{\overline{V}wN_{l_j}/K} \geq e^{w^2N_{l_j}/K}) \\ &\leq \tfrac{1}{e^{w^2N_{l_j}/K}}E[e^{\overline{V}wN_{l_j}/K}] \\ &= \tfrac{1}{e^{w^2N_{l_j}/K}}E[e^{(V_1+V_2+\ldots+V_{N_{l_j}})w/K}] \\ &= \tfrac{1}{e^{w^2N_{l_j}/K}}\textstyle\prod_{i=1}^{N_{l_j}}\psi_i(\tfrac{w}{K}) \\ &\leq \tfrac{1}{e^{w^2N_{l_j}/K}}e^{w^2N_{l_j}/(8K)} \\ &= e^{-7w^2N_{l_j}/(8K)}. \end{aligned}$$

Let p_N denote the empirical distribution of X_n defined as $p_N(l_j) = \frac{1}{N}\{n : 1 \leq n \leq N$ and $X_n \in l_j\}$. By Assumption 4, we have $p_N(l_j) \geq k_N\frac{logN}{N}$, where k_N are a sequence of positive constants and $\lim_N k_N = \infty$, we have

$$\begin{aligned} Pr(\overline{V} \geq w) &\leq e^{-7w^2 N_{l_j}/(8K)} \\ &= e^{-7w^2 p_N(l_j)N/(8K)} \\ &\leq e^{-7w^2 k_N logN/(8K)} \\ &= N^{-7w^2 k_N/(8K)} \\ &= N^{-7w^2 k_N/[8(b-a)^2]}. \end{aligned}$$

Similarly, we have $Pr(\overline{V} \leq -w) \leq N^{-7w^2 k_N/[8(b-a)^2]}$. Therefore, we have $Pr(|\overline{V}| \geq w) \leq 2N^{-7w^2 k_N/[8(b-a)^2]}$.

B.2.3 Proof of Theorem 1

To prove Theorem 1, we use a fundamental combinatorial result due to Vapnik and Chervonenkis (1971). Since each group in π is a convex polyhedron in the d-dimensional Euclidean space with at most M faces, there exists a collection ϕ of the subsets of the set $\{X_1, X_2, ..., X_n\}$ such that $\#(\phi) \leq (2N)^{M(d+2)}$ and ϕ has the property that, for any polyhedron j with no more than M faces in the d-dimensional Euclidean space, there exists an $s \in \phi$ such that $i \in l_j$ if and only if $X_i \in s$. Note that, for subgroup l_j, we have $\hat{\beta}_{IV}(x_{l_j}) - \beta_{l_j} = (\frac{1}{N_{l_j}} \sum_{i \in l_j} Z_i T_i)^{-1} (\frac{1}{N_{l_j}} \sum_{i \in l_j} Z_i \epsilon_i) = Q_{l_j}^{-1} \overline{V}$ (see e.g., Greene 2003, Chapter 5). These imply that, for any $w > 0$, we have

$$\begin{aligned} & \lim_{N \to \infty} Pr(\max_j |\hat{\beta}_{IV}(x_{l_j}) - \beta_{l_j}| \geq w) \\ = & \lim_{N \to \infty} Pr(\bigcup_j \{|\hat{\beta}_{IV}(x_{l_j}) - \beta_{l_j}| \geq w\}) \\ = & \lim_{N \to \infty} Pr(\bigcup_j \{|Q_{l_j}^{-1} \overline{V}| \geq w\}) \\ \leq & \lim_{N \to \infty} (\min_j |Q_{l_j}|)^{-1} \times (2N)^{M(d+2)} \times 2N^{-7w^2 k_N/[8(b-a)^2]} \\ = & Q^{*-1} \lim_{N \to \infty} \times 2^{1+M(d+2)} N^{M(d+2)-7w^2 k_N/[8(b-a)^2]} \\ = & 0. \end{aligned}$$

where $Q^* = min_j |Q_{l_j}|$. The second inequality follows from Lemma 2 and the fundamental combinatorial result; the third equality follows from Assumption 3, which states that Q is bounded for all subgroups; and the last equality follows from $\lim_N k_N = \infty$ (Assumption 4). Because $\beta_i \xrightarrow{p} \beta_{l_j}$ (from Lemma 1), we have

$$
\begin{aligned}
& \lim_{N\to\infty} Pr(\max_j |\hat{\beta}_{IV}(x_{l_j}) - \beta_i| \geq w) \\
= & \lim_{N\to\infty} Pr(\max_j |\hat{\beta}_{IV}(x_{l_j}) - \beta_{l_j} + \beta_{l_j} - \beta_i| \geq w) \\
\leq & \lim_{N\to\infty} Pr(\max_j (|\hat{\beta}_{IV}(x_{l_j}) - \beta_{l_j}| + |\beta_{l_j} - \beta_i|) \geq w) \\
\leq & \lim_{N\to\infty} Pr(\max_j |\hat{\beta}_{IV}(x_{l_j}) - \beta_{l_j}| \geq \tfrac{w}{2}) \;\; or \lim_{N\to\infty} Pr(\max_j |\beta_{l_j} - \beta_i| \geq \tfrac{w}{2}) \\
\leq & \lim_{N\to\infty} Pr(\max_j |\hat{\beta}_{IV}(x_{l_j}) - \beta_{l_j}| \geq \tfrac{w}{2}) + \lim_{N\to\infty} Pr(\max_j |\beta_{l_j} - \beta_i| \geq \tfrac{w}{2}) \\
= & \; 0.
\end{aligned}
$$

B.2.4 Proof of Corollary 1

To prove that the causal tree does not provide a consistent estimator, it suffices to show that there exists l_j such that $\hat{\beta}_{CT}(x_{l_j}) \overset{p}{\nrightarrow} \beta_{l_j}$.

$$
\begin{aligned}
\hat{\beta}_{CT}(x_{l_j}) &= E(Y_{1i}|T_i = 1) - E(Y_{2i}|T_i = 0) \\
&= \beta_{l_j} + [E(\varepsilon_i|T_i = 1) - E(\varepsilon_i|T_i = 0)].
\end{aligned}
$$

Let p denote the probability of receiving a treatment, i.e., $p = P(T_i = 1)$. By expanding $Cov(T_i, \varepsilon_i)$, we have

$$
\begin{aligned}
Cov(T_i, \varepsilon_i) &= E(T_i\varepsilon_i) - E(T_i)E(\varepsilon_i) \\
&= pE(\varepsilon_i|T_i = 1) - [pE(\varepsilon_i|T_i = 1) + (1-p)E(\varepsilon_i|T_i = 0)]p \\
&= p(1-p)[E(\varepsilon_i|T_i = 1) - E(\varepsilon_i|T_i = 0)].
\end{aligned}
$$

Because $Cov(T_i, \varepsilon_i) \neq 0$ implies that $E(\varepsilon_i|T_i = 1) - E(\varepsilon_i|T_i = 0) \neq 0$, we have $\hat{\beta}_{CT}(x_{l_j}) \neq \beta_{l_j}$.

APPENDIX C

Appendix to Chapter 4

C.1 Evaluation of the Instrument

To check if travel distance correlates with patient sickness, we analyze if patients living closer to a hospital are sicker or healthier as indicated by their age, number of chronic conditions and number of comorbidities. The results are summarized below. We do not see evidence of such correlation.

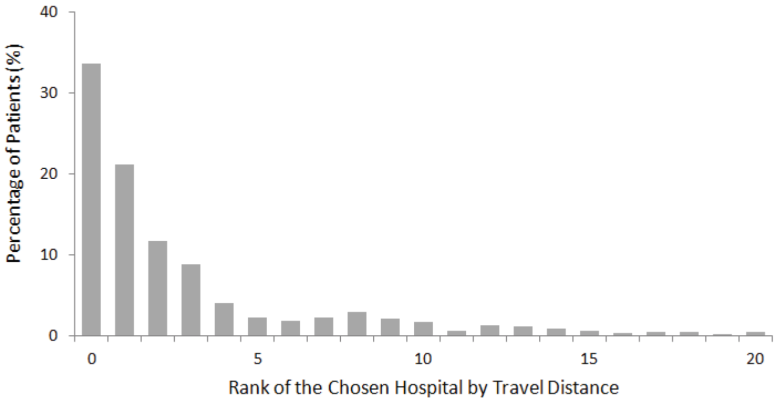
Table C.1: Relationship between Distance to the Nearest Hospital and Patient Characteristics

Distance (in mile)	Number of Patients	Patients' Mean Age	Number of Chronic Conditions	Number of Comorbidities
below 5	47,192	67.6 (12.3)	6.6 (2.4)	2.7 (1.5)
5 to 10	18,015	68.7 (11.8)	6.7 (2.4)	2.7 (1.5)
above 10	32,978	67.6 (11.8)	6.6 (2.5)	2.6 (1.5)
Total	98,185	67.8 (12.1)	6.6 (2.4)	2.7 (1.5)

To provide a sense of the correlation between travel distance and patient choice, we first rank hospitals by travel distance for each patient and then analyze the overall ranks of the chosen hospitals. Figure C.1 shows the percentage of patients and the ranks of their chosen hospitals. We see that more than half of the patients chose the nearest or the second nearest hospitals. We also see that the probability of choosing a hospital decreases as the distance increases.

We check the strength of the instrumental variable by regressing the provider indicator over the instrumental variable for each group of patients (i.e., first stage) when we compare each pair of providers. The coefficients are significant at 1% significance level in 99% of the cases. These results suggest that the instrumental variable has a strong first stage.

Figure C.1: Rank of the Chosen Hospital by Travel Distance



BIBLIOGRAPHY

Ang, E., Kwasnick, S., Bayati, M., Plambeck, E. L., Aratow, M. (2015). Accurate emergency department wait time prediction. Manufacturing & Service Operations Management, 18(1), 141-156.

Athey, S., Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. Proceedings of the National Academy of Sciences, 113(27), 7353-7360.

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. The Annals of Statistics, 47(2), 1148-1178.

Ayanian, J. Z., Weissman, J. S. (2002). Teaching hospitals and quality of care: a review of the literature. The Milbank Quarterly, 80(3), 569-593.

Ban, G. Y., El Karoui, N., & Lim, A. E. (2016). Machine learning and portfolio optimization. Management Science, 64(3), 1136-1154.

Barr, J. K., Giannotti, T. E., Sofaer, S., Duquette, C. E., Waters, W. J., Petrillo, M. K. (2006). Using public reports of patient satisfaction for hospital quality improvement. Health Services Research, 41(3p1), 663-682.

Barro, J. R., Huckman, R. S., Kessler, D. P. (2006). The effects of cardiac specialty hospitals on the cost and quality of medical care. Journal of Health Economics, 25(4), 702-721.

Bartel, A. P., Chan, C. W., Kim, H. (2016). Should hospitals keep their patients longer?: The role of inpatient and outpatient care in reducing readmissions. National Bureau of Economic Research.

Bastani, H., Bayati, M. (2016). Online decision-making with high-dimensional covariates. Working paper, Graduate School of Business, Stanford University, Stanford, CA.

Bastani, H., Goh, J., Bayati, M. (2018). Evidence of upcoding in pay-for-performance programs. Management Science. https://doi.org/10.1287/mnsc.2017.2996. [Epub ahead of print].

Batt, R. J., Terwiesch, C. (2015). Waiting patiently: An empirical study of queue abandonment in an emergency department. Management Science, 61(1), 39-59.

Bavafa, H., Hitt, L. M., & Terwiesch, C. (2018). The impact of e-Visits on visit frequencies and patient health: Evidence from primary care. Management Science, 64(12), 5461-5480.

Bertsimas, D., O'Hair, A., Relyea, S., Silberholz, J. (2016). An analytics approach to designing combination chemotherapy regimens for cancer. Management Science, 62(5), 1511-1531.

Birkmeyer, J. D., Siewers, A. E., Finlayson, E. V., Stukel, T. A., Lucas, F. L., Batista, I., ..., Wennberg, D. E. (2002). Hospital volume and surgical mortality in the United States. New England Journal of Medicine, 346(15), 1128-1137.

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

Black, W. C., Gareen, I. F., Soneji, S. S., Sicks, J. D., Keeler, E. B., Aberle, D. R., ..., Gatsonis, C. (2014). Cost-effectiveness of CT screening in the National Lung Screening Trial. New England Journal of Medicine, 371(19), 1793-1802.

Bock, R. D., Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. Psychometrika, 46(4), 443-459.

Bolling, S. F., Li, S., O'Brien, S. M., Brennan, J. M., Prager, R. L., Gammie, J. S. (2010). Predictors of mitral valve repair: clinical and surgeon factors. The Annals of Thoracic Surgery, 90(6), 1904-1912.

Boscoe, F. P., Henry, K. A., Zdeb, M. S. (2012). A nationwide comparison of driving distance versus straight-line distance to hospitals. The Professional Georgrapher, 64(2), 188-196.

Bourguignon, T., Bouquiaux-Stablo, A. L., Loardi, C., Mirza, A., Candolfi, P., Marchand, M., Aupart, M. R. (2014). Very late outcomes for mitral valve replacement with the Carpentier-Edwards pericardial bioprosthesis: 25-year follow-up of 450 implantations. The Journal of Thoracic and Cardiovascular Surgery, 148(5), 2004-2011.

Breen, R., Choi, S., Holm, A. (2015). Heterogeneous causal effects and sample selection bias. Sociological Science, 2, 351-369.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Breiman, L., Friedman, J., Stone, C. J., Olshen, R. A. (1984). Classification and regression trees. CRC press.

Brooks, J. M., Irwin, C. P., Hunsicker, L. G., Flanigan, M. J., Chrischilles, E. A., Pendergast, J. F. (2006). Effect of dialysis center profit status on patient survival: A comparison of risk - adjustment and instrumental variable approaches. Health Services Research, 41(6), 2267-2289.

Burge, P., Devlin, N., Appleby, J., Gallo, F., Nason, E., Ling, T. (2005). A model of patients' choices of hospital from stated and revealed preference choice data. The London Patient Choice Project Team, Department of Health London. The RAND Corporation.

Carroll, R. J., Horn, S. D., Soderfeldt, B., James, B. C., Malmberg, L. (1995). International comparison of waiting times for selected cardiovascular procedures. Journal of the American College of Cardiology, 25(3), 557-563.

Chan, C. W., Farias, V. F., Escobar, G. J. (2016). The impact of delays on service times in the intensive care unit. Management Science, 63(7), 2049-2072.

Chassin, M. R., Kosecoff, J., Park, R. E., Winslow, C. M., Kahn, K. L., Merrick, N. J., ..., Brook, R. H. (1987). Does inappropriate use explain geographic variations in the use of health care services? A study of three procedures. JAMA, 258(18), 2533-2537.

Chipman, H. A., George, E. I., McCulloch, R. E. (2010). BART: Bayesian additive regression trees. The Annals of Applied Statistics, 4(1), 266-298.

Clark, J. R., Huckman, R. S. (2012). Broadening focus: Spillovers, complementarities, and specialization in the hospital industry. Management Science, 58(4), 708-722.

Clark, J. R., Huckman, R. S., Staats, B. R. (2013). Learning from customers: Individual and organizational effects in outsourced radiological services. Organization Science, 24(5), 1539-1557.

Cox, C. E., Carson, S. S., Govert, J. A., Chelluri, L., Sanders, G. D. (2007). An economic evaluation of prolonged mechanical ventilation. Critical Care Medicine, 35(8), 1918.

Daneshmand, M. A, Milano, C. a, Rankin, J.S., et al. (2009). Mitral valve repair for degenerative disease: A 20-year experience. The Annals of Thoracic Surgery, 88(6), 1828-37.

Daneshmand, M. A., Milano, C. A., Rankin, J. S., Honeycutt, E. F., Shaw, L. K., Davis, R. D., ..., Smith, P. K. (2010). Influence of patient age on procedural selection in mitral valve surgery. The Annals of Thoracic Surgery, 90(5), 1479-1486.

Dixon, A., Robertson, R., Appleby, J., Burge, P., Devlin, N. J. (2010). Patient choice: how patients choose and how providers respond. King's Fund.

Duggirala, A. V., Chen, F. M., Gergen, P. J. (2004). Postoperative adverse events in teaching and nonteaching hospitals. Family Medicine, 36(7), 508-513.

Elixhauser, A., Steiner, C., Harris, D. R., Coffey, R. M. (1998). Comorbidity measures for use with administrative data. Medical Care, 36(1), 8-27.

Emmert, M., Schlesinger, M. (2016). Hospital Quality Reporting in the United States: Does Report Card Design and Incorporation of Patient Narrative Comments Affect Hospital Choice?. Health Services Research, 52(3), 933-958.

Fasken, L. L., Wipke-Tevis, D. D., Sagehorn, K. K. (2001). Factors associated with unplanned readmissions following cardiac surgery. Progress in Cardiovascular Nursing, 16(3), 107-115.

Ferreira, K. J., Lee, B. H. A., Simchi-Levi, D. (2015). Analytics for an online retailer: Demand forecasting and price optimization. Manufacturing & Service Operations Management, 18(1), 69-88.

Fineberg, S. J., Oglesby, M., Patel, A. A., Pelton, M. A., Singh, K. (2013). Outcomes of cervical spine surgery in teaching and non-teaching hospitals. Spine, 38(13), 1089-1096.

Finlayson, S., Birkmeyer, J., Tosteson, A., Nease, R. (1999). Patient preferences for location of care, implications for regionalization. Medical Care, 37(2), 204-209.

Freeman, M., Savva, N., Scholtes, S. (2016). Gatekeepers at work: An empirical analysis of a maternity unit. Management Science, 63(10), 3147-3167.

Gammie, J. S., Sheng, S., Griffith, B. P., Peterson, E. D., Rankin, J. S., O'Brien, S. M., Brown, J. M. (2009). Trends in mitral valve surgery in the United States: Results from the Society of Thoracic Surgeons Adult Cardiac Surgery Database. The Annals of Thoracic Surgery, 87(5), 1431-1437.

Gaynor, M., Propper, C., Seiler, S. (2010, February). The Effect of Patient Choice: Evidence from Recent NHS Reforms.

Gelsomino, S., Lorusso, R., Livi, U., Masullo, G., Lucà, F., Maessen, J., Gensini, G. F. (2011). Cost and cost-effectiveness of cardiac surgery in elderly patients. The Journal of Thoracic and Cardiovascular Surgery, 142(5), 1062-1073.

Gerteis, M. (1993). Through the patient's eyes: Understanding and promoting patient-centered care. San Francisco: Jossey-Bass.

Gibbons, R. D., Bock, R. D. (1987). Trend in correlated proportions. Psychometrika, 52(1), 113-124.

Gibbons, R. D., Hedeker, D. (1997). Random effects probit and logistic regression models for three-level data. Biometrics, 53(4), 1527-1537.

Glance, L. G., Osler, T. M., Mukamel, D. B., Dick, A. W. (2007). Effect of complications on mortality after coronary artery bypass grafting surgery: Evidence from New York State. The Journal of Thoracic and Cardiovascular Surgery, 134(1), 53-58.

Gopaldas, R. R., Bakaeen, F. G., Dao, T. K., Coselli, J. S., LeMaire, S. A., Huh, J., Chu, D. (2012). Outcomes of concomitant aortic valve replacement and coronary artery bypass grafting at teaching hospitals versus nonteaching hospitals. The Journal of Thoracic and Cardiovascular Surgery, 143(3), 648-655.

Gowrisankaran, G., & Town, R. J. (1999). Estimating the quality of care in hospitals using instrumental variables. Journal of Health Economics, 18(6), 747-767.

Green, D. P., Kern, H. L. (2012). Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees. Public Opinion Quarterly, 76(3), 491-511.

Greene, W. H. (2003). Econometric analysis, 5th edn. Prentice Hall, Upper Saddle River.

Groux, P., Anchisi, S., Szucs, T. (2014). Are cancer patients willing to travel more or further away for a slightly more efficient therapy? Cancer and Clinical Oncology, 3(1), 36-42.

Guajardo, J. A., Cohen, M. A., Netessine, S. (2015). Service competition and product quality in the US automobile industry. Management Science, 62(7), 1860-1877.

Gupta, P. K., Fernandes-Taylor, S., Ramanan, B., Engelbert, T. L., Kent, K. C. (2014). Unplanned readmissions after vascular surgery. Journal of Vascular Surgery, 59(2), 473-482.

Hahn, P. R., Murray, J., & Carvalho, C. M. (2017). Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects. Confounding, and Heterogeneous Effects. Working paper, School of Mathematics and Statistical Sciences, Arizona State University, Tempe, Arizona.

Hausman, J. A. (1978). Specification tests in econometrics. Econometrica: Journal of the Econometric Society, 46(6), 1251-1271.

Heller, R., Rosenbaum, P. R., Small, D. S. (2009). Split samples and design sensitivity in observational studies. Journal of the American Statistical Association, 104(487), 1090-1101.

Ho, V., Hamilton, B. H., & Roos, L. L. (2000). Multiple approaches to assessing the effects of delays for hip fracture patients in the United States and Canada. Health Services Research, 34(7), 1499-1518.

Hothorn, T., Hornik, K., Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. Journal of Computational and Graphical Statistics, 15(3), 651-674.

Huckman, R. S., Kelley, M. A. (2013). Public reporting, consumerism, and patient empowerment. New England Journal of Medicine, 369(20), 1875-1877.

Huckman, R. S., Pisano, G. P. (2006). The firm specificity of individual performance: Evidence from cardiac surgery. Management Science, 52(4), 473-488.

Huckman, R. S., & Zinner, D. E. (2008). Does focus improve operational performance? Lessons from the management of clinical trials. Strategic Management Journal, 29(2), 173-193.

Hutton, D. W., Brandeau, M. L., So, S. K. (2011). Doing good with good OR: supporting cost-effective hepatitis B interventions. Interfaces, 41(3), 289-300.

Hwang, J. S., Tsauo, J. Y., Wang, J. D. (1996). Estimation of expected quality adjusted survival by cross-sectional survey. Statistics in Medicine, 15(1), 93-102.

Imai, K., Ratkovic, M. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. The Annals of Applied Statistics, 7(1), 443-470.

Institute of Medicine (US). Committee on Quality of Health Care in America. (2001). Crossing the quality chasm: A new health system for the 21st century. National Academy Press.

Iribarne, A., Chang, H., Alexander, J. H., et al. (2014). Readmissions after cardiac surgery: Experience of the NIH/CIHR cardiothoracic surgical trials network. The Annuals of Thoracic Surgery, 98(4), 1274-1280.

Jaeker, J., Tucker, A. (2016). Past the point of speeding up: The negative effects of workload saturation on efficiency and quality. Management Science, 63(4), 1042-1062.

Jidéus, L., Liss, A., Ståhle, E. (2009). Patients with sternal wound infection after cardiac surgery do not improve their quality of life. Scandinavian Cardiovascular Journal, 43(3), 194-200.

Kang, J. H., Chen, Y. H., Lin, H. C. (2010). Comorbidity profiles among patients with ankylosing spondylitis: a nationwide population-based study. Annals of the Rheumatic Diseases, Annals of Rheumaic Diseases, 69(6), 1165-1168.

Kattan, M. W., Vickers, A. J. (2004). Incorporating predictions of individual patient risk in clinical trials. In Urologic Oncology: Seminars and Original Investigations, (22)4, 348-352.

KC, D., Staats, B. R. (2012). Accumulating a portfolio of experience: The effect of focal and related experience on surgeon performance. Manufacturing & Service Operations Management, 14(4), 618-633.

KC, D., Staats, B. R., Gino, F. (2013). Learning from my success and from others' failure: Evidence from minimally invasive cardiac surgery. Management Science, 59(11), 2435-2449.

KC, D., Terwiesch, C. (2011). The effects of focus on performance: Evidence from California hospitals. Management Science, 57(11), 1897-1912.

KC, D., Terwiesch, C. (2012). An econometric analysis of patient flows in the cardiac intensive care unit. Manufacturing & Service Operations Management, 14(1), 50-65.

Keeler, E. B., Rubenstein, L. V., Kahn, K. L., Draper, D., Harrison, E. R., McGinty, M. J., ... , Brook, R. H. (1992). Hospital characteristics and quality of care. Jama, 268(13), 1709-1714.

Kent, D. M., Hayward, R. A. (2007). Limitations of applying summary results of clinical trials to individual patients: The need for risk stratification. JAMA, 298(10), 1209-1212.

Khuri, S. F., Najjar, S. F., Daley, J., Krasnicka, B., Hossain, M., Henderson, W. G., ... DePalma, R. (2001). Comparison of surgical outcomes between teaching and non-teaching hospitals in the Department of Veterans Affairs. Annals of Surgery, 234(3), 370.

Kilic, A., Shah, A. S., Conte, J. V., Baumgartner, W. A., Yuh, D. D. (2013). Operative outcomes in mitral valve surgery: Combined effect of surgeon and hospital volume in a population-based analysis. The Journal of Thoracic and Cardiovascular Surgery, 146(3), 638-646.

Kim, S. H., Chan, C. W., Olivares, M., Escobar, G. (2014). ICU admission control: An empirical study of capacity allocation and its implication for patient outcomes. Management Science, 61(1), 19-38.

King, G., & Nielsen, R. (2016). Why propensity scores should not be used for matching. Copy at http://j. mp/1sexgVw Download Citation BibTex Tagged XML Download Paper, 378.

Kravitz, R. L., Duan, N., Braslow, J. (2004). Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. Milbank Quarterly, 82(4), 661-687.

Lacy, A. M., García-Valdecasas, J. C., Delgado, S., Castells, A., Taurá, P., Piqué, J. M., & Visa, J. (2002). Laparoscopy-assisted colectomy versus open colectomy for treatment of non-metastatic colon cancer: a randomised trial. The Lancet, 359(9325), 2224-2229.

LaPar, D. J., Hennessy, S., Fonner, E., Kern, J. a, Kron, I. L., Ailawadi, G. (2010). Does urgent or emergent status influence choice in mitral valve operations? An analysis of outcomes from the Virginia Cardiac Surgery Quality Initiative. The Annals of Thoracic Surgery, 90(1), 153-60.

Lee, D. D., Li, J., Wang, G., Croome, K. P., Burns, J. M., Perry, D. K., ... & Taner, C. B. (2017). Looking inward: The impact of operative time on graft survival after liver transplantation. Surgery, 162(4), 937-949.

Lieberman, J. A., Stroup, T. S., McEvoy, J. P., Swartz, M. S., Rosenheck, R. A., Perkins, D. O., ... & Severe, J. (2005). Effectiveness of antipsychotic drugs in patients with chronic schizophrenia. New England Journal of Medicine, 2005(353), 1209-1223.

Lindenauer, P. K., Remus, D., Roman, S., Rothberg, M. B., Benjamin, E. M., Ma, A., Bratzler, D. W. (2007). Public reporting and pay for performance in hospital quality improvement. New England Journal of Medicine, 356(5), 486-496.

Lingsma, H. F., Steyerberg, E. W., Eijkemans, M. J. C., Dippel, D. W. J., Scholte Op Reimer, W. J. M., Van Houwelingen, H. C. (2010). Comparing and ranking hospitals based on outcome: results from The Netherlands Stroke Survey. QJM: An International Journal of Medicine, 103(2), 99-108.

Listl, S., Jürges, H., & Watt, R. G. (2016). Causal inference from observational data. Community Dentistry and Oral Epidemiology, 44(5), 409-415.

Lu, S. F., Lu, L. X. (2016). Do mandatory overtime laws improve quality? Staffing decisions and operational flexibility of nursing homes. Management Science, 63(11), 3566-3585.

Lu, S. F., Rui, H., Seidmann, A. (2017). Does technology substitute for nurses? Staffing decisions in nursing homes. Management Science. https://doi.org/10.1287/mnsc.2016.2695. [Epub ahead of print].

Lu, Y., Musalem, A., Olivares, M., Schilkrut, A. (2013). Measuring the effect of queues on customer purchases. Management Science, 59(8), 1743-1763.

Mant, D. (1999). Can randomised trials inform clinical decisions about individual patients? The Lancet, 353(9154), 743-746.

Masoomi, H., Wirth, G. A., Paydar, K. Z., Richland, B. K., Evans, G. R. (2014). Perioperative outcomes of autologous breast reconstruction surgery in teaching versus nonteaching hospitals. Plastic and Reconstructive Surgery, 134(4), 514e-520e.

McClellan, M., McNeil, B. J., Newhouse, J. P. (1994). Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality: analysis using instrumental variables. Jama, 272(11), 859-866.

Menendez, M. E., Neuhaus, V., van Dijk, C. N., Ring, D. (2014). The Elixhauser comorbidity method outperforms the Charlson index in predicting inpatient death after orthopaedic surgery. Clinical Orthopaedics and Related Research, 472(9), 2878-2886.

Merkow, R. P., Ju, M. H., Chung, J. W., Hall, B. L., Cohen, M. E., Williams, M. V., ..., Bilimoria, K. Y. (2015). Underlying reasons associated with hospital readmission following surgery in the United States. JAMA, 313(5), 483-495.

Moran, J. L., Solomon, P. J., ANZICS Centre for Outcome and Resource Evaluation (CORE) of the Australian, New Zealand Intensive Care Society (ANZICS. (2014). Fixed effects modelling for provider mortality outcomes: Analysis of the Australia and New Zealand Intensive Care Society (ANZICS) adult patient data-base. PloS one, 9(7), e102297.

Moss, R. R., Humphries, K. H., Gao, M., Thompson, C. R., Abel, J. G., Fradet, G., & Munt, B. I. (2003). Outcome of mitral valve repair or replacement: A comparison by propensity score analysis. Circulation, 108(10 suppl 1), II-90.

Nandyala, S. V., Marquez-Lara, A., Fineberg, S. J., Hassanzadeh, H., & Singh, K. (2014). Complications after lumbar spine surgery between teaching and nonteaching hospitals. Spine, 39(5), 417-423.

Nilsdotter, A., Bremander, A. (2011). Measures of hip function and symptoms: Harris Hip Score (HHS), Hip Disability and Osteoarthritis Outcome Score (HOOS), Oxford Hip Score (OHS), Lequesne Index of Severity for Osteoarthritis of the Hip (LISOH), and American Academy of Orthopedic Surgeons (AAOS) Hip and Knee Questionnaire. Arthritis Care & Research, 63(S11).

Paddock, S. M., Adams, J. L., de la Guardia, F. H. (2015). Better-than-average and worse-than-average hospitals may not significantly differ from average hospitals: an analysis of Medicare Hospital Compare ratings. BMJ Quality and Safety, 24(2), 128-134.

Porembka, M. R., Hall, B. L., Hirbe, M., & Strasberg, S. M. Quantitative weighting of postoperative complications based on the accordion severity grading system: Demonstration of potential impact using the american college of surgeons national surgical quality improvement program. Journal of the American College of Surgeons, 210(3), 286-298.

Powers, S., Qian, J., Jung, K., Schuler, A., Shah, N. H., Hastie, T., Tibshirani, R. (2017). Some methods for heterogeneous treatment effect estimation in high-dimensions. Working paper, Stanford Graduate School of Business, Stanford University, Stanford, CA.

Ramdas, K., Saleh, K., Stern, S., Liu, H. (2017). Variety and experience: Learning and forgetting in the use of surgical devices. Management Science. https://doi.org/10.1287/mnsc.2016.2721. [Epub ahead of print].

Ray, S., Chambers, J., Gohlke-Baerwolf, C., Bridgewater, B. (2006). Mitral valve repair for severe mitral regurgitation: the way forward? European Heart Journal, 27(24), 2925-2928.

Reddy, H. G., Shih, T., Englesbe, M. J., Shannon, F. L., Theurer, P. F., Herbert, M. A., ... Prager, R. L. (2013). Analyzing "failure to rescue": is this an opportunity for outcome improvement in cardiac surgery? The Annals of Thoracic Surgery, 95(6), 1976-1981.

Regier, D. A., Sunderji, R., Lynd, L. D., Gin, K., Marra, C. A. (2006). Cost-effectiveness of self-managed versus physician-managed oral anticoagulation therapy. Canadian Medical Association Journal, 174(13), 1847-1852.

Ruel, M., Rubens, F. D., Masters, R. G., Pipe, A. L., Bédard, P., Mesana, T. G. (2004). Late incidence and predictors of persistent or recurrent heart failure in patients with mitral prosthetic valves. The Journal of Thoracic and Cardiovascular Surgery, 128(2), 278-283.

Russo, A., Grigioni, F., Avierinos. J-F, et al. (2008). Thromboembolic complications after surgical correction of mitral regurgitation incidence, predictors, and clinical implications. Journal of the American College of Cardiology, 51(12), 1203-11.

Savage, E. B., Ferguson, T. B., DiSesa, V. J. (2003). Use of mitral valve repair: Analysis of contemporary United States experience reported to the Society of Thoracic Surgeons National Cardiac Database. The Annals of Thoracic Surgery, 75(3), 820-5.

Schneider, C. R., Cobb, W., Patel, S., Cull, D., Anna, C., Roettger, R. (2009). Elective surgery in patients with end stage renal disease: what's the risk? The American Surgeon, 75(9), 790-793.

Shah, S. V., Gage, B. F. (2011). Cost-effectiveness of dabigatran for stroke prophylaxis in atrial fibrillation clinical perspective. Circulation, 123(22), 2562-2570.

Signorovitch, J. E. (2007). Identifying informative biological markers in high-dimensional genomic data and clinical trials. (Doctoral dissertation, Harvard University).

Silverstein, M. D., Qin, H., Mercer, S. Q., Fong, J., Haydar, Z. (2008). Risk factors for 30-day hospital readmission in patients $\geq$ 65 years of age. Proceedings (Baylor University. Medical Center), 21(4), 363-372.

Sinaiko, A. D., Eastman, D., Rosenthal, M. B. (2012). How report cards on physicians, physician groups, and hospitals can have greater impact on consumer choices. Health Affairs, 31(3), 602-611.

Society of Thoracic Surgeons (2016). Online STS risk calculator.

Song, H., Tucker, A. L., Murrell, K. L. (2015). The diseconomies of queue pooling: An empirical investigation of emergency department length of stay. Management Science, 61(12), 3032-3053.

Strasberg, S. M., Hall, B. L. (2011). Postoperative morbidity index: a quantitative measure of severity of postoperative complications. Journal of the American College of Surgeons, 213(5), 616-626.

Su, X., Tsai, C. L., Wang, H., Nickerson, D. M., Li, B. (2009). Subgroup analysis via recursive partitioning. Journal of Machine Learning Research, 10(Feb), 141-158.

Sullivan, P. W., Ghushchyan, V. (2006). Preference-based EQ-5D index scores for chronic conditions in the United States. Medical Decision Making, 26(4), 410-420.

Taddy, M., Gardner, M., Chen, L., & Draper, D. (2016). A nonparametric bayesian analysis of heterogenous treatment effects in digital experimentation. Journal of Business & Economic Statistics, 34(4), 661-672.

Taylor Jr, D. H., Whellan, D. J., Sloan, F. A. (1999). Effects of admission to a teaching hospital on the cost and quality of care for Medicare beneficiaries. New England Journal of Medicine, 340(4), 293-299.

Tian, L., Alizadeh, A. A., Gentles, A. J., Tibshirani, R. (2014). A simple method for estimating interactions between a treatment and a large number of covariates. Journal of the American Statistical Association, 109(508), 1517-1532.

Thornlow, D. K., Stukenborg, G. J. (2006). The association between hospital characteristics and rates of preventable complications and adverse events. Medical Care, 44(3), 265-269.

Tsai, T. C., Jha, A. K., Gawande, A. A., Huckman, R. S., Bloom, N., Sadun, R. (2015). Hospital board and management practices are strongly related to hospital performance on clinical quality metrics. Health Affairs, 34(8), 1304-1311.

US Food and Drug Administration (2013). Paving the way for personlized medicine: FDA's role in the new era of medical product development.

van Tuinen, M., Elder, S., Link, C., Li, S., Song, J. H., Pritchett, T. (2005). Surveillance of surgery-related adverse events in Missouri using ICD-9-CM codes.

van Walraven, C., Austin, P. C., Jennings, A., Quan, H., & Forster, A. J. (2009). A modification of the Elixhauser comorbidity measures into a point system for hospital death using administrative data. Medical Care, 47(6), 626-633.

Vapnik, V. N. and A. Y., Chervonenkis (1971). Theory of Probability and Its Applications, 16, 264.

Vartak, S., Ward, M. M., Vaughn, T. E. (2008). Do postoperative complications vary by hospital teaching status? Medical Care, 46(1), 25-32.

Vassileva, C. M., Shabosky, J., Boley, T., Markwell, S., Hazelrigg, S. (2012). Cost analysis of isolated mitral valve surgery in the United States. The Annals of Thoracic Surgery, 94(5), 1429-1436.

Vassileva, C. M., Mishkel, G., McNeely, C., Boley, T., Markwell, S., Scaife, S., Hazelrigg, S. (2013). Long-term survival of patients undergoing mitral valve repair and replacement: A longitudinal analysis of Medicare fee-for-service beneficiaries. Circulation, 127(18), 1870-6.

Vassileva, C. M., Boley, T., Standard, J., Markwell, S., Hazelrigg, S. (2013). Relationship between patient income level and mitral valve repair utilization. Heart Surgery Forum, 16(2), E89-95.

Velentgas, P., Dreyer, N. A., Nourjah, P., Smith, S. R., & Torchia, M. M. (Eds.). (2013). Developing a protocol for observational comparative effectiveness research: a user's guide. Government Printing Office.

Vuik, S. I., Mayer, E. K., & Darzi, A. (2016). Patient segmentation analysis offers significant benefits for integrated care and support. Health Affairs, 35(5), 769-775.

Wager, S., Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association, 113(523), 1228-1242.

Wang, G., Li, J., Hopp, W., Fazzalari, F., Bolling, S. (2018). Cost-effectiveness of referring patients to centers of excellence for mitral valve surgery. Working Paper. Ross School of Business, University of Michigan, Ann Arbor, MI.

Williams, J. F., MoRRow, A. G., Braunwald, E. (1965). The incidence and management of "medical" complications following cardiac operations. Circulation, 32(4), 608-619.

Windisch, W., Freidel, K., Schucher, B., Baumann, H., Wiebel, M., Matthys, H., Petermann, F. (2003). Evaluation of health-related quality of life using the MOS 36-Item Short-Form Health Status Survey in patients receiving noninvasive positive pressure ventilation. Intensive Care Medicine, 29(4), 615-621.

Wooldridge, J. M. (2010). Econometric analysis of cross section and panel data. MIT Press, Cambridge, MA.

World Health Organization. (2011). Global atlas on cardiovascular disease prevention and control. World Health Organization in collaboration with the World Heart Federation and the World Stroke Organization.

Xie, Y., Brand, J. E., Jann, B. (2012). Estimating heterogeneous treatment effects with observational data. Sociological methodology, 42(1), 314-347.

Xu, Y., Armony, M., Ghose, A. (2017). The Interplay between online reviews and physician demand: An empirical investigation. Working paper, Gies College of Business, University of Illinois, Urbana-Champaign, IL.

Zaric, G. S., Brandeau, M. L., Barnett, P. G. (2000). Methadone maintenance and HIV prevention: a cost-effectiveness analysis. Management Science, 46(8), 1013-1031.

Zeileis, A., Hothorn, T., Hornik, K. (2008). Model-based recursive partitioning. Journal of Computational and Graphical Statistics, 17(2), 492-514.