ESSAYS ON PAYMENT REFORM MODELS AND CAPACITY PLANNING IN HEALTHCARE

A Dissertation

by

SEOKJUN YOUN

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Chelliah Sriskandarajah
Committee Members,	H. Neil Geismar
	Gregory R. Heim
	Subodha Kumar
	Andrew L. Johnson
Head of Department,	Richard Metters

May 2019

Major Subject: Business Administration

ABSTRACT

My dissertation is inspired by challenging yet encouraging payment policies and operational issues in the U.S. healthcare system. Two of the three essays delivers policy implications for bundled payment reform models that aim to improve the predictability of care outcomes and associated costs by reducing variation in care-delivery practices. To investigate how variation in practice relates to hospital operational performance, my first essay proposes a novel measure of clinical practice variation based on a high-volume inpatient discharge dataset. I find that hospitals may be improperly rewarded for quality improvements if practice variation is ignored, implying that incentives and penalties for hospital operations should be designed to account for such effects.

Also, I identify potential drawbacks inherent in the government's status quo policy for selecting participating providers in the bundled payment reform models. To address this issue, my second essay incorporates insights from the first essay and suggests a systematic framework for healthcare provider evaluation and selection. Using a combinatorial auction model equipped with data envelopment analysis as a pre-selection tool, the proposed framework alleviates the inherent decision-making bias of the current system and deploys adequate healthcare providers for target regions, thereby creating an optimized bundled payment program.

Lastly, my third essay applies a process improvement perspective to study adaptive capacity planning in ambulatory surgery centers. Timely capacity adjustment is essential for the surgery center planners as each facility is concerned with the cost and capacity implications of adding/removing specific surgical procedures under the transition toward payment reform models. But, the related research is limited. In contrast to the traditional top-down approach to capacity planning, my approach proposes a bottom-up strategy based on optimization methods combined with analytics that are informed by operational-level archival patient flow data. I develop several mathematical formulations and heuristics based on scheduling theory to derive the most cost-efficient capacity solution for the multi-stage structure of surgery centers. In the computational study, I further show how uncertain business parameters may affect capacity planning decisions.

DEDICATION

To my father Seungwon, my mother Kwangyuon, my sister Hyunjin,

and my dear wife Yookyung.

Thank you for all of your support along the way.

ACKNOWLEDGEMENTS

I would never have been able to finish my dissertation without the guidance of my committee members and coauthors, help from friends, and support from my family and my wife.

I would like to express my deepest gratitude to my advisor Dr. Chelliah Sriskandarajah, for his excellent supervision, guidance, caring, and patience throughout my doctoral education. Simply put, his mentorship helped me not only to develop the right set of skills to succeed in the academic field but also to stay strong and to be patient while going through hard times. I hope I can be an insightful and generous mentor like Dr. Chelliah Sriskandarajah in the future.

I wish to thank Dr. Neil Geismar, Dr. Gregory Heim, and Dr. Subodha Kumar, for their guidance through research projects and for their service as members in my doctoral committee. They are certainly great mentors and co-authors to work with. I would also like to thank the rest of my committee members and coauthors: Dr. Anupam Agrawal, Dr. Andrew Johnson, and Dr. Vikram Tiwari for their encouragement and insightful comments on my dissertation.

I want to express my sincere gratitude to Dr. Rich Metters for his leadership as the Department Chair and for his efforts to make the life of doctoral students more enjoyable. Furthermore, I also want to thank Dr. James Abbey, Dr. Andres Jola-Sanchez, Dr. Michael Ketzenberg, Dr. Rogelio Oliva, Dr. Bala Shetty, and Dr. Jon Stauffer for their excellent guidance on my job search, as well as Ms. Donna Shumaker, Ms. Tammy Louther, Ms. Veronica Stilley, and the rest of the department for the help I got so freely from them.

To my PhD mates in the department, thanks for being so helpful and constructive. Special thanks go to Seung Jun Lee, Serkan Akturk, Rakesh Mallipeddi, Xingzhi Jia, and Yating Feng for being supportive and encouraging friends.

Lastly, I would like to express my deep gratitude to my parents, who have taught me the core values of virtue, dignity, honesty, and respect. My father, mother, and sister have always been a source of inspiration and motivation for me. I always know and feel that they are with me and that they never let me feel alone. Special thanks to my wife for her never ending help and moral support.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supervised by a dissertation committee consisting of Professors Chelliah Sriskandarajah [advisor], H. Neil Geismar, and Gregory R. Heim of the Department of Information and Operations Management, Professor Andrew L. Johnson of the Department of Industrial and Systems Engineering at Texas A&M University, and Professor Subodha Kumar of the Department of Marketing and Supply Chain Management at Temple University.

The data analyzed for Chapter 3 was provided by Professor Vikram Tiwari at the Vanderbilt University Medical Center. All other work conducted for the dissertation was completed by the student under the guidance of the dissertation committee members.

Funding Sources

Graduate study was supported by fellowships from Mays Business School and the Department of Information and Operations Management at Texas A&M University. Chapter 3 was supported by a research grant from Mays Business School at Texas A&M University.

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

High costs and inconsistent care-quality are long-lasting issues in the U.S. healthcare system (OECD, 2016). Most experts indicate the current dominant payment system called fee-for-service is one of the critical sources of the problems (Hussey et al., 2011), because it incentivizes volume rather than a quality of care. As a remedy, several payment reform policies have been initiated to fix the incentive structure and ultimately to improve the suboptimal healthcare system (Green, 2012).

In particular, my dissertation is inspired by the bundled payment reform policy that aims to improve the predictability of care outcomes and associated costs by reducing variation in care-delivery practices. Specifically, the bundled payment policy combines a number of care-delivery services needed for an episode of patient care and develops a contractual agreement between a payer and a care provider, such that the reimbursement is fixed *a priori* regardless of eventual cost. As such, the bundled payment schemes are meant to reduce healthcare costs and cost variations by establishing transparent and standardized clinical pathways (Abecassis, 2015). However, because the healthcare industry must provide services to individual patients having unique characteristics and symptoms, whether such a bundled payment standardization strategy can be effective in reducing the cost of care episode while improving the care quality is not clear.

In this regard, my first essay delivers *policy implications* for the bundled payment models by devising a novel measure of clinical practice variation and further examining how variation in practice relates to hospital operational performance. Also, I identify potential drawbacks inherent in the government's status quo policy for selecting participating providers in the nationwide bundled payment models. To support a smooth transition toward the reform models, my second essay incorporates insights from the first essay and suggests a systematic *policy implementation* framework designed for healthcare provider evaluation and selection. Meanwhile, such a transition is worthwhile if healthcare processes also promote efficient performance. For example, timely capacity adjustment is essential for ambulatory surgery centers as each facility is concerned with the cost and capacity implications of adding/removing specific surgical procedures under the transition toward payment reform models. Hence, my third essay applies a *process improvement* perspective to study adaptive capacity planning in the surgery centers, where the problem presents challenges, yet the related research is few. In what follows, I briefly describe each essay in sequence.

1.1 Policy Implications: Impacts of Clinical Practice Variation on Operational Performance

The first essay explores whether and how smaller variations in clinical practice improve hospital operational performance. The extant literature discusses geographical variations in healthcare (Ham, 1988; Clancy and Cronin, 2005) but often ignores clinical variations that may be prevalent in a hospital. To fill this research gap, I propose a stepwise approach to measure practice variation defined as all variation not resulting from the patient-mix, and I indeed observe a broad practice variation spectrum across hospitals. Using statistical process control as a theoretical lens (Oakland, 2007), I hypothesize that such practice variation adversely affects operational performance. I analyze six years of a high-volume inpatient discharge dataset from New York and Florida, and after accounting for the dynamic endogeneity of practice variation and operational performance (Arellano and Bover, 1995; Blundell and Bond, 1998), I find supporting evidence that greater practice variation is associated with longer patient stays and higher total cost per capita. Interestingly, this phenomenon is even stronger when a hospital provides a higher-quality patient experience because such a hospital tends to provide more responsive care, which is often resource-intensive. Therefore, hospitals may be improperly rewarded for quality improvements if practice variation is ignored, implying that incentives and penalties for hospital operations should be designed to account for such effects.

To suggest actionable improvement plans, I also delve into the granular level of practice variation, including the risk associated with under-ordering laboratory/radiology tests. Such test-ordering practices are generally conducted in an early stage of the entire care episode (Alexander, 2012; Zhi et al., 2013). Hence, the underuse of test-ordering practices may lead to extra efforts in care-delivery stages to alleviate potential negative effects of the test underuse. I find that higher underuse variation in the test-ordering may lead to higher care-delivery cost, again affected by quality evaluations. Hence, too restrictive reimbursement rules on test-ordering practice are undesirous, as they may result in even higher eventual total care-delivery spending. Regarding the causes of practice variation, I further find that organizational environments (Westert and Groenewegen, 1999), rather than physicians' personal behaviors (Wennberg and Gittelsohn, 1973), possibly drive the level of practice variation, implying the opportunity of hospital performance improvements via well-designed bundled payment policies. This study contributes by highlighting the importance of measuring and understanding practice variation and suggests how managers and policy-makers can use the findings to design better bundled payment reform models.

1.2 Policy Implementation: Healthcare Provider Evaluation and Selection Framework

Given the heterogeneity of clinical practice and performance across hospitals, as I study in the first essay, it is critical that bundled payment reform identifies competent healthcare providers. To address this issue, my second essay proposes a practical and systematic provider selection framework from the perspective of a payer that operates a bundled payment program. Currently, the government creates a weighted average composite score from a multitude of dimensions related to a hospital's characteristics and performance, and it uses this score to select participating providers for the bundled payment program (CMS, 2014c). Despite the popularity of the weighted score approach in practice, however, biases and additive assumptions introduced in the development of the weights and the evaluation can cause the inherent problems in the decision-making process (McCabe et al., 2005).

My proposed selection framework mitigates the potential drawbacks of the existing approach by using a combinatorial auction model equipped with data envelopment analysis as a pre-selection tool. Specifically, the framework contributes by proposing an approach to sequentially compare separate dimensions of participating providers rather than merely comparing a composite score while allowing flexibility for each provider in selecting the bundle of services. A numerical study supports that, beyond alleviating the potential drawbacks of the current system, the proposed framework deploys adequate healthcare providers for target regions, thereby creating an optimized bundled payment program. Hence, I carefully point out that the lack of systematic provider selection is possibly one reason why the ongoing nationwide bundled payment initiative does not reveal apparent simultaneous achievements in cost reduction and care quality improvement (CMS, 2015a, 2016a). To address practical issues in this framework, I have communicated with the Center for Medicare & Medicaid Innovation to collect application data on the ongoing nationwide bundled payment initiative for future extension of this study. The data collection is approved via the Freedom of Information Act.

1.3 Process Improvement: Adaptive Capacity Planning for Ambulatory Surgery Centers

The explosion of patient data is changing the way and the extent to which healthcare organizations capture data, analyze information, and make decisions. In my third essay, I use patient flow data to develop novel models that drive capacity decisions. Specifically, I propose an adaptive capacity planning model for ambulatory surgery centers (ASCs) that have transformed the outpatient experience for millions of people by offering a convenient, personalized, lower-priced alternative to hospitals.

Planning ASC capacity is challenging especially due to the multi-stage nature of surgical services and significant uncertainty in the patient-mix and service durations of patients at each stage (Tiwari and Sandberg, 2016). Furthermore, the ASC planners face with the dynamic environment of operations, e.g., Medicare frequently updates the list of surgical procedures allowable in ASCs because of several reasons such as reimbursement policy adjustments and technology advancements. Extant literature suggests efficient patient scheduling decision for established facilities (Ruiz and Vázquez-Rodríguez, 2010; Ribas et al., 2010), but there is limited research that provides capacity planning/renovation tools under such dynamic environments. The third essay contributes by filling this research gap in the scheduling and healthcare resource allocation literature. A notable trade-off in this process is between the need to be responsive to patients' demand by having enough capacity and the need to efficiently schedule surgeries to increase the capacity utilization by minimizing the costs related to overtime of the ASC resources.

In contrast to the traditional top-down approach to capacity planning (Hans et al., 2012), my approach proposes a bottom-up strategy based on optimization methods combined with analytics benefiting from the detailed operational-level patient flow data. To be more specific, I use analytics tools to classify patients into a few groups, which reduces the complexity of the capacity planning problem and improves the model's practicality. Later, I develop several mathematical formulations and heuristics based on scheduling theory to derive the most cost-efficient capacity solution for the multi-stage structure of surgery centers. I explicitly consider three sequential stages that are typical patient flows in ASCs: the pre-operative stage at the holding room, the intra-operative stage at the operating room, and the post-operative stage at the post-anesthesia care unit. In the computational study, I further show how uncertain business parameters may affect capacity planning decisions to provide managerial implications for the practitioners. This study highlights the benefit of considering multiple stages together in capacity planning, rather than focusing on the largest cost center (e.g., the operating rooms as in Mancilla and Storer (2012); Mak et al. (2014)) exclusively. This research will be impactful in guiding more than 5,000 ASCs in the U.S., performing 23 million surgeries annually (MedPAC, 2017), to make appropriate investments that will improve ASC operations on capacity adjustment and patient scheduling.

In summary, my dissertation develops managerial and theoretical contributions in healthcare operations supported by grounded theory from practitioners' insights and literature in statistical process control theory, economic theory, and scheduling theory. Based on the three essays in the domain of healthcare operations management, I provide a research portfolio that contributes to each area of policy implications, policy implementation, and process improvement by using both empirical analyses and analytical modeling. The following chapters demonstrate timely and relevant issues for both practitioners and academic researchers in the healthcare industry.

The remainder of this dissertation is structured as follows. Chapter 2 examines the impacts of clinical practice variation on operational performance to provide implications for bundled payment reform models. Chapter 3 develops a healthcare provider evaluation and selection framework designed for the bundled payment models. Chapter 4 devises an adaptive capacity planning approach for healthcare providers with a multi-stage structure in the context of ambulatory surgery centers. Chapter 5 summarizes contributions and concludes this dissertation.

2. EXAMINING IMPACTS OF CLINICAL PRACTICE VARIATION ON OPERATIONAL PERFORMANCE: IMPLICATIONS FOR BUNDLED PAYMENT REFORM MODELS

"The intensifying pressure from employers and insurers for transparent pricing is already beginning to force providers to explain – or eliminate – hard-to-justify price variations" -Porter and Lee (p. 59, 2013)

2.1 Introduction

Due to excessive national expenditures, lack of transparent pricing, and a history of often poor health outcomes in the U.S. healthcare system (Mossialos et al., 2016; OECD, 2016), policy-makers are experimenting with various payment reform models (Green, 2012; Vanberkel et al., 2012). These payment reform models attempt to improve quality of care and to reduce healthcare spending (Hussey et al., 2011). For example, bundled payment models combine a number of care-delivery services needed for an episode of patient care, and develop a contractual agreement between a payer and a care provider, to fix reimbursement a priori regardless of eventual cost. As such, bundled payment models are meant to deliver appropriate care yet reduce healthcare costs and cost uncertainty for patients, employers who pay for their insurance, and third-party payers.

The envisioned payment reform models often aim concomitantly to decrease variation in care-delivery practices, resulting in more predictable care outcomes and costs (Abecassis, 2015). For example, providers operating under bundled payment programs are forced to move toward lower practice variation via standardized care bundles. Yet, because the healthcare industry must provide customized services to individual patients having unique characteristics and symptoms, whether such a standardization strategy can be effective in achieving its main goals (i.e., improving care quality while reducing episode treatment cost) is not clear.

Motivated by such objectives, this study aims to empirically examine links between the level of practice variation inside a hospital and its operational performance (measured as patient length-of-stay and care-delivery cost). A distinct challenge with understanding potential impacts of payment reform models concerns a total lack of publicly available historical data about most such programs, including contractual bundled payments. We circumvent this challenge by recasting available fee-for-service (FFS) data into relevant metrics to address how reducing practice variation, such as via bundled payment contracts, might impact hospital performance. Considering the high costs and inconsistent care quality of the U.S. healthcare system in an environment of limited resources (Shi and Singh, 2012), it is societally meaningful to investigate whether and how lessening variations in clinical practice might relate to reform model goals.

2.1.1 Research Questions

A major research opportunity exists in precisely measuring practice variation observed during the process of delivering healthcare, and identifying its impacts on operational performance. Existing literature classifies practice variations into warranted and unwarranted (Clancy and Cronin, 2005; Appleby et al., 2011). Case-mix-index (CMI)[1] is often used to reflect variation due to a hospital's patient mix, which we will call warranted variation. CMI controls for diversity, clinical complexity, and resource needs of the patient population in a hospital, enabling fair hospital-to-hospital comparisons of medical operations. However, certain care delivery variations exist even after considering a hospital-level patient mix, thus CMI cannot capture all aspects of variation. Wennberg (2002) called all variation not resulting from patient mix by the term unwarranted variation. Unwarranted variation can result from variation in care providers' decisions, variation to customize care, or variation in medical procedure supplies. Because prior work focuses on CMI, the impact of these unwarranted variations on hospital performance is not clear. Surprisingly, however, no CMI-like nationally standardized metric for unwarranted variations exists. We thus identify a major research opportunity for measuring unwarranted-type variations in medical practice for patient cohorts having the identical condition (hereafter, we will use the term *practice variation* to refer to the overall contribution of these three unwarranted variation facets). Subsequently, to bridge this research gap, our first research question is: *(1) How can researchers measure practice variation for a patient cohort with reasonable precision?*

[1] A list of acronyms in this study is provided in Appendix A.1.

Another objective is to study the potential impacts on hospital performance of payment reform models that implicitly pursue, yet do not explicitly mandate, low practice variation. To the best of our knowledge, however, only a few hospitals have fully adopted new initiatives such as bundled payment models (Tsai et al., 2015) and The Centers for Medicare & Medicaid Services (CMS) is not making data publicly available on experiences under those models. An experiment might be an alternative option to compare such performance, by randomly assigning hospitals to adopt specific payment reform models, but this approach is ethically problematic. Instead, we envision an opportunity to measure the existing practice variation and compare hospitals using readily available FFS data. We assume hospitals that have low practice variation will resemble hospitals operating under a payment reform model. Thus, we close the research gap by asking: *(2) Does lower practice variation of a hospital (or within patient cohort care episode groups) necessarily relate to better operational performance?* Relying on statistical process control (SPC) theory (Wheeler et al., 1992; Oakland, 2007), we posit that high practice variation may harm hospital operational performance.

Besides reducing practice variation, payment reform models often incorporate several quality initiatives to encourage better care quality. We believe these quality initiatives may change participating healthcare providers' behaviors. Thus, we also consider possible effects of these quality initiatives on the link between practice variation and hospital performance. Process quality and experiential quality are two key measures of care quality that are publicly reported, span salient aspects of operations, and are used by CMS to evaluate healthcare providers (Sadeghi et al., 2012). *Process Quality* concerns how well a hospital adheres to evidence-based medical guidelines to diagnose and treat patients (Theokary and Ren, 2011; Chandrasekaran et al., 2012; Nair et al., 2013; Andritsos and Tang, 2014b). *Experiential Quality* aggregates the reports of patients about their observations of and participation in healthcare (Sadeghi et al., 2012), and thus relates to external perceptions of care quality from a patient's perspective (Donabedian, 1980; Li and Benton, 1996). Both process quality and experiential quality are relevant exogenously mandated measures that may influence effectiveness of payment reform models. Yet, whether these two quality metrics play effective roles in the relationship between practice variation and operational performance is unexplored.

Analysis of the roles of these two quality metrics on the link between practice variation and performance is important since a tension may occur when hospitals aim to improve both process quality and experiential quality simultaneously. Chandrasekaran et al. (2012) conceptually connect process quality and experiential quality with exploitation (that reduces variation) and exploration (that induces variation), respectively, two types of organizational learning activities in tension with each other (March, 1991; Benner and Tushman, 2003). We also surmise that hospitals' efforts to improve the two quality measures can affect a hospital's operational actions. We thus consider practice variation together with the two quality metrics. Following literature in Total Quality Management (Evans and Lindsay, 1999; Douglas and Judge, 2001), process quality may alleviate negative effects of practice variation on operational performance. In contrast, experiential quality, which is designed to fulfill specific subjective requirements of patients (Nair et al., 2013), may lead to responsive patient treatments that perhaps deviate from guidelines, leading to an amplified negative relationship between practice variation and operational performance.[2]

With the growing impetus for well-designed payment reform models and the potentially confounding influence of mandated quality measurement initiatives, we investigate our next research question: *(3) Does the level of quality measures affect the relationship between practice variation and operational performance?* Our study contributes by addressing how empirical evidence about the effect of process quality and experiential quality on operational performance can change if we explicitly consider practice variation inside a hospital. This issue merits investigation since the joint consideration of practice variation and mandated quality measures may offer directions for quality improvement initiatives and payment reform models.

Next, we delve into granular levels of practice variation to investigate whether practice variation in hospital test-ordering can differently affect performance within subsequent care-delivery stages. We focus on underuse practice variation in the test-ordering stage because (a) most patients (i.e.,

[2] This line of reasoning does not intend to ignore the value of experiential quality that is examined in the operations management (OM) literature (e.g., Nair et al., 2013; Senot et al., 2015), but to shed light on potential detrimental effects of experiential quality on operational performance when considered jointly with practice variation.

about 98%) go through at least one laboratory or radiology test, (b) on average the test-ordering accounts for a large proportion (about 27%) of total charges per patient,[3] and more importantly (c) test-ordering practices are generally conducted in an early stage of the entire care episode (Zhi et al., 2013; Alexander, 2012).

We posit that higher underuse practice variation in test-ordering can lead to higher subsequent care-delivery costs resulting from extra efforts to alleviate negative effects of the test underuse. In healthcare, the growing realization about the importance of preventive care can be understood in a similar vein because many experts believe preventive care is indeed more cost-effective than reactive care (Nutting, 1994; Fox and Shaw, 2015). Accordingly, we look at our next research question: *(4) Does underuse practice variation in test-ordering differently relate to operational performance in the care-delivery stage?* Answering this question will help shed light on the impact of serial practice variations, which is understudied in the healthcare industry, and will provide insights for designing policy incentives.

2.1.2 Key Findings and Contributions

We analyze six years of a comprehensive dataset from 387 hospitals in NY and FL states, focusing on Medicare patients with three conditions: acute myocardial infarction (AMI), heart failure (HF), and pneumonia (PN). To examine hypotheses, we use a dynamic panel model with System Generalized Method of Moments (GMM) estimation (Arellano and Bover, 1995; Blundell and Bond, 1998) to account for the dynamic endogeneity of practice variation and operational performance. Using several new metrics to measure practice variation, we observe that the level of practice variation varies across hospitals even when they have the same level of CMI. We first find that higher overall hospital practice variation relates to significantly longer average patient length-of-stay (LOS), but rather weakly higher average total cost per patient. We delve into variation in test-ordering practice, to shed light on a possible trade-off between test-ordering and care-delivery practices that compromise total treatment cost. Interestingly, we find that hospitals with a higher risk of underusing tests may face unexpected higher expenditures in subsequent

[3] See Appendix A.8 for details.

care-delivery stages. These findings are even stronger if the hospital provides service having high experiential quality, because such a hospital pays attention to patient experience and thus tries to mitigate potential negative consequences by devoting more time and resources to these custom experiences.

Overall, our findings deliver important implications for payment reform programs that are fundamentally designed to promote reduced practice variation. For example, under bundled payments, a patient may stay a shorter period in a hospital during her entire episode of care. However, the cost savings from the program might be uncertain. Bundling all services for an episode of care (i.e., providing a standard set of care as much as possible) may not be the best approach. Instead, we suggest policy-makers should allow some flexible practice or lenient reimbursement standards, perhaps similar to the experimental bundled payment schemes that rely on retrospective fee-for-service based reimbursement. According to our study findings, this approach may be particularly effective for the laboratory/radiology test-ordering practice. By establishing flexible reimbursement standards for test-ordering practices, possibly with an upper-limit cap, health providers may not only reduce failures or underuse of meaningful tests but also may alleviate the risk of higher care-delivery spending, as our findings imply. By doing so, the healthcare system may ultimately find an opportunity to reduce total cost per patient. Our measure of practice variation also contributes since it enables researchers and managers to rigorously measure and visualize the status of hospitals' practice variation.

The remainder of the paper is organized as follows. Section 2 provides background and develops hypotheses. Section 3 presents data, variables of interest, research methods, and econometric models, followed by estimation results in section 4. Section 5 provides a broader discussion of our findings and implications. Finally, section 6 concludes and provides directions for future research.

2.2 Background and Hypotheses Development

In this section, we first contrast fee-for-service healthcare delivery against bundled payment reform model characteristics. We then review literature on clinical practice variation as well as process and experiential quality. Finally, we develop hypotheses pertaining to practice variation in hospitals.

2.2.1 Bundled Payment Program

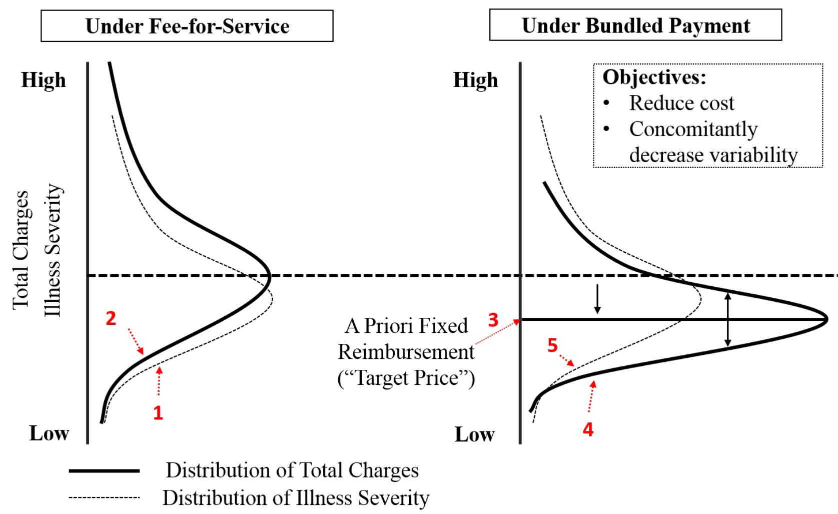
Traditionally, the U.S. federal government Medicare insurance program makes separate payments to healthcare providers for each of the individual services they furnish to beneficiaries for a single course of treatments. This approach, called fee-for-service (FFS), mis-incentivizes providers to perform more treatments and to use more equipment regardless of their necessity for a care episode (Arrow et al., 2009). Such outcomes reflect practice variation. FFS thus is believed to be one of the main causes of high healthcare spending (Hussey et al., 2011).

As an alternative, policy-makers suggest the use of payment reform models such as bundled payment (BP) contracts, in which an *a priori* fixed reimbursement between payers and providers applies for an entire episode of patient care (Miller, 2009). CMS claims bundled payment models may spur hospitals, physicians, and other providers to better coordinate care, improve quality of care, and consider the financial implications of their decisions, leading to lower spending (CMS, 2015a). The U.S. Department of Health and Human Services (HHS) aims to tie 50% of traditional FFS Medicare payments to care quality through Accountable Care Organizations (ACOs) or bundled payment arrangements by the end of 2018 (HHS, 2015). This aim leads top management of healthcare providers to consider whether to participate in these programs.

To illustrate potential implications of the move from FFS toward payment reform models, Figure 2.1 compares the distribution of patient illness severity and total charges between FFS and BP. Under FFS, total charges that reflect clinical practice (Line 2) tend to be escalated compared to the actual illness severity of patients (Line 1). BP proposes an *a priori* fixed reimbursement (Line 3), which is set up as lower than the average historical total charge under FFS, incentivizing providers to eliminate unnecessary spending (Tsai et al., 2015). Under BP, if a hospital's spending for a patient is more than the fixed BP reimbursement, the hospital is responsible for the portion of over-charge. Meanwhile, due to care quality requirements in BP, hospitals cannot merely provide minimum care to ensure a patient's profitability. Thus, under BP the variance of the total charge distribution also should become smaller than that of FFS, reflecting lower practice variation (Line

4). The challenging part is that the exogenous patient population distribution perhaps remains the same as with FFS (Line 5).

Figure 2.1: Illustration: FFS vs. Bundled Payment



In the OM literature, limited analytical work compares BP schemes to FFS along outcomes such as extent of patient selection, treatment intensity, financial risk (Adida et al., 2016), patient welfare, waiting time (Guo et al., 2016), and readmission rate (Andritsos and Tang, 2018). Gupta and Mehrotra (2015) discuss provider selection and information sharing under a stylized BP system. Most of these papers however solely examine patient variation, assuming that treatment is given to patients proportionally to their level of illness, while ignoring providers' levels of practice variation, which is our contribution.

Does reducing practice variation help to achieve care quality and cost reduction objectives, or improve operational performance? We find an opportunity to examine this question by comparing magnitudes of practice variation across hospitals under FFS, to examine how reduced practice variation tends to relate to operational performance. In effect, most of the bundled reform program initiatives that have been widely applied in practice are “retrospective” payment models still operating on a FFS payment scheme (CMS, 2015a), meaning that hospitals file claims with a payer (either a commercial payer or CMS) based on the agreed upon FFS schedule; afterwards, the payer performs an adjudication of the expenses related to the condition at the basis of the care-episode bundle.[4] Hence, our analysis using FFS inpatient data is a valid approach to evaluate/monitor the practice variation of healthcare providers under such schemes.

2.2.2 Clinical Practice Variation

Due to persistent healthcare variation (Ham, 1988; Clancy and Cronin, 2005), practitioners and researchers have tried to identify where the variation originates (Miller et al., 2011), to disentangle warranted variation from unwarranted variation (Wennberg, 2002), and to investigate how unwarranted variation can be addressed (Appleby et al., 2011). Wennberg et al. (2002) divide unwarranted variation into three categories: effective care, preference-sensitive care, and supply-sensitive care. Effective care refers to care delivery services of proven effectiveness, thus variation will reflect failure to deliver needed care. Preference-sensitive care involves care decisions based on patient preferences and values. Supply-sensitive care refers to the frequency of resource usage (e.g., equipment for laboratory and radiology tests) that is governed by the decisions of a care provider. These three care categories reflect the varying decision-making processes under differing clinical theory, medical evidence, patient preferences, and supply of resources (Sipkoff, 2003).

Measuring variation for the three separate care categories, as they are defined above, is challenging. For example, since every patient has different types and levels of conditions, the proper amount and frequency of needed care becomes a subjective decision made by a physician in coordination with the patient and other stakeholders. As such, without differentiating those three categories, we aim to capture the level of practice variation within a patient care episode cohort. Because every hospital has its own "chargemaster" and captures charges for all services and items provided (Melnick and Fonkych, 2008; Ferenc, 2013), we use detailed medical charge information to construct a measure of practice variation. Our study contributes by constructing a precise measure of practice variation applicable to both hospital and condition (i.e., care episode) level. In

[4] If the sum of the claims is below the agreed-upon price, hospitals share savings. Otherwise, any overage is incurred by the payer under upside-only BP design.

contrast to prior work that examines hospital-level geographic variations, our work tracking the extent of practice variation inside a hospital may facilitate development of incentives that encourage actions to deal with practice variation.

2.2.3 Process Quality and Experiential Quality

Due to preventable errors in U.S. hospitals, many patients die or are injured in a year (Kohn et al., 2000). To avoid such errors and to evaluate healthcare quality, the Hospital Inpatient Quality Reporting program, launched in 2003, mandates the collection and disclosure of process quality measures, for example, the percentage of patients who receive treatments known to lead to the best results (CMS, 2014b). Hospitals are motivated to report their process quality to receive financial incentives (CMS, 2010). Prior operations management literature sheds light on the impacts of higher process quality under various research settings. On one hand, Nair et al. (2013) find that higher process quality and clinical flexibility reduce patient LOS in a cardiology unit. Andritsos and Tang (2014b) subdivide process quality into clinical and administrative dimensions, and find that higher clinical process quality is associated with lower patient LOS. On the other hand, medical experts point out that documentation and monitoring of process quality are resource-intensive tasks (Fonarow and Peterson, 2009). Improving process quality can incur substantial costs for hospitals (Senot et al., 2015). Collectively, prior literature identifies inconsistent associations between process quality and resource usage, for which LOS and cost are often used as proxy measures.

While standardized clinical processes may improve operational performance, they may not address the specific requirements of patients (Nair et al., 2013). Indeed, patients also can evaluate perceived quality attributes, such as experiential quality. The Hospital Consumer Assessment of Healthcare Providers and Systems (HCHAPS) survey measures include communication with caregivers and their responsiveness to patients' requests. Consideration of experiential quality is important because caregivers (e.g., physicians and nurses) can use communication to learn more about patients, come up with an accurate diagnosis, and select the most appropriate care method depending on the individual patient's characteristics (Elwyn et al., 2000). Increased attention to each patient's characteristics has been shown to associate with patient satisfaction (Rubin et al., 2001), lower readmission rates (Senot et al., 2015), and decreased patient LOS (Nair et al., 2013). On the other hand, being responsive to patients or communicating more frequently with them can require substantial resources such as advanced information systems (Khunlertkit and Carayon, 2013), rapid response teams (Kapu et al., 2014), and responsive registered nurses (Smolowitz et al., 2015), leading to higher cost per patient discharge (Senot et al., 2015).

Taken together, previous studies view process quality and experiential quality as either complementary assets (Toussaint, 2009; Nair et al., 2013) or trade-offs (Chandrasekaran et al., 2012) in achieving various operating outcomes or performance metrics. As far as we know, however, our work is the first study that examines how the relationship between the two quality dimensions and operational performance changes once a hospital's practice variation is considered.

2.2.4 Hypotheses Development

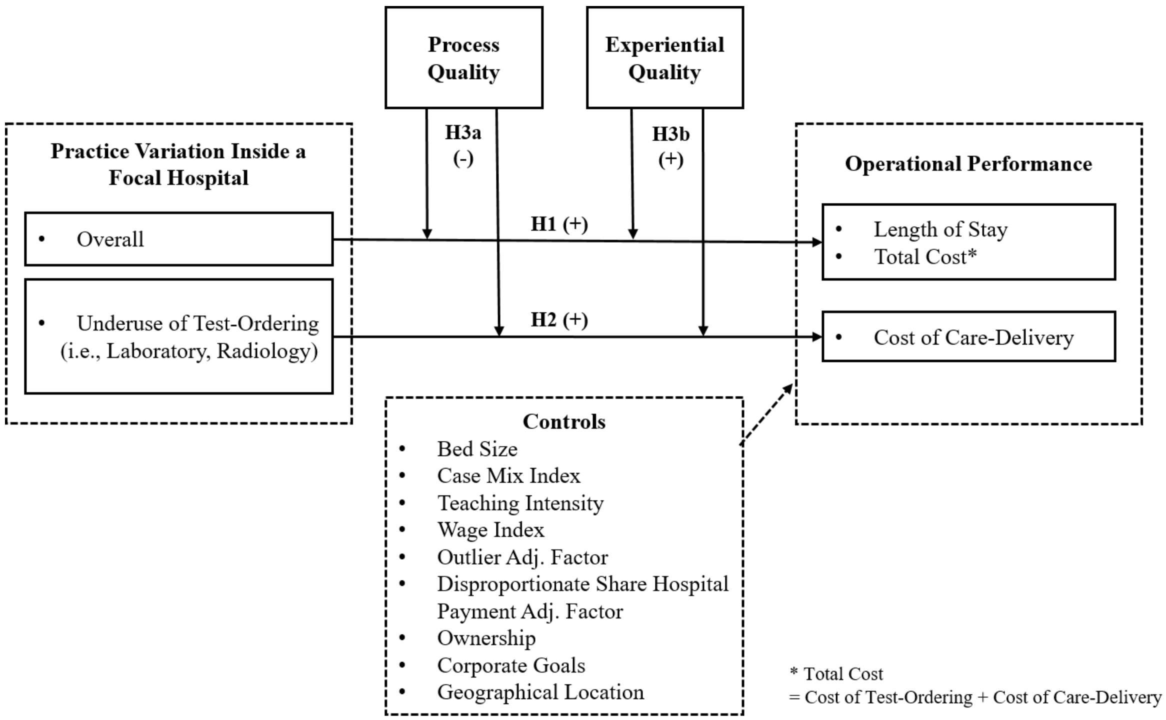
We hypothesize a relationship between practice variation and resource usage (Hypothesis 1). We delve into practice variation, focusing on underuse risk of test-ordering, to see its impact on subsequent care-delivery cost (Hypothesis 2). We also motivate how process quality (Hypothesis 3a) and experiential quality (Hypothesis 3b) differently influence these relationships. Our conceptual framework is shown in Figure 2.2.

2.2.4.1 Impact of Practice Variation on Patient LOS and Total Cost per Capita

For reimbursement purposes, hospitals record all expenditures for services and equipment administered to each patient. Hence, patient-level detailed medical charge information captures the clinical pathway that the patient went through. As such, variations in medical charges provide an opportunity to measure variability in the underlying treatment process (Melnick and Fonkych, 2008; Ferenc, 2013). If considerable practice variation exists in the claim charges within a hospital for patients having the same medical condition, managers need to examine where the variation originates and how it can be addressed to meet payment reform model goals (Appleby et al., 2011).

Quality management literature distinguishes between two types of variation that can cause quality problems (Garvin, 1988; Cachon and Terwiesch, 2008). One type is *common variation*,

Figure 2.2: Conceptual Framework



purely random variations in an output of a process. The other type is *assignable variation*, whose source can be identified and possibly managed. Key objectives of quality management are to ensure that the outputs of a process are consistent (i.e., the process is in control) and meet the customer's expectations (i.e., the process is capable). If the process is out of control or not capable, quality management teams must identify assignable variation and reduce it. We conceptually link the assignable variation concept to practice variation.

Considering the potential adverse effect of assignable variation on output, we posit that high practice variation may harm hospital-level care delivery efficiency. Using statistical process control (SPC) to monitor and control a process enables it to reach its full potential (Wheeler et al., 1992; Oakland, 2007), in that process managers can minimize waste and produce as much conforming output as possible. Similarly, in the healthcare context, assembling all the necessary medical services for a common care episode (e.g., AMI, HF, or PN) is a recurring managerial decision process. After a patient is admitted to a hospital, the sequence of recurring decisions made by the administrative group, physicians, and nurses will be reflected in the charges for each patient.

Thus, unstable charges for patients with the same disease and similar illness severity indicate unstable care processes. This process instability possibly increases variability in service time during a care episode, which is known to cause waiting times (Cachon and Terwiesch, 2008; Gupta et al., 2016). Using a queueing framework, for example, Dai et al. (2016) claim that physicians' decisions about ordering tests influence patients' waiting times, which may drive overall service time, or patient length-of-stay. Unstable hospital care processes might also increase the uncertainty of patient outcomes, possibly leading hospitals to devote additional efforts/resources to alleviate negative outcomes that may critically harm the hospital's reputation.

Collectively, variable service times may lead to longer patient LOS, whether due to value-added time when providers are delivering appropriate care, or non-value added time when a patient is staying without receiving care. Similarly, any chance of uncertain outcomes may lead hospitals to incur a greater amount of expenditure for patient treatment, whether due to meaningful efforts when providers are delivering appropriate care, or due to wasteful practice when a patient is given excessive care. Lower practice variation may reduce resource usage for patient care by eliminating non-value added waiting, lessening over-treatment, limiting excessive patient customization, and lowering material supply uncertainty. Hence, we hypothesize a positive relationship between practice variation and resource usage, proxied as patient LOS and average total cost per patient.

Hypothesis 1. *Higher clinical practice variation is associated with greater resource usage, for which per patient LOS and total cost per capita are used as proxy measures.*

2.2.4.2 Impact of Variation in Test-Ordering Practice on Care-Delivery Cost

While we first focus on practice variation impacts on LOS and total cost during the entire process of a patient's care episode, we recognize that each care episode comprises several serial stages, potentially including admission, diagnosis, treatment, recovery, and discharge. By splitting the entire care episode into test-ordering stages and care-delivery stages (although they are not perfectly sequential and often go back and forth), we recognize potential impacts of practice variation during the test-ordering stages on the average cost for the care-delivery stages. From SPC theory, detecting problems at an early stage and preventing them from occurring is ideal for any organization (Wheeler et al., 1992). In this sense, we focus on variation in test-ordering practice and provide our theoretical reasoning for its possible impact on cost.

Many existing explanations imply that variations in medical practice are caused by different physicians' preferences about appropriate treatment for patients (i.e., practice style), often called the practice style hypothesis (Wennberg and Gittelsohn, 1973; Stano, 1993). From this perspective, physicians differ in the kind of procedures they provide because they learned to value them differently, and thus adhere to a different practice style. The difference in practice style may emerge because of uncertainty about the value of a certain medical procedure (Wennberg and Gittelsohn, 1982). Uncertainty also is often viewed as a personal attribute, for example, physicians have different tolerance toward uncertainty. Thus, if physicians are uncertain about how to proceed in a given case, they may try to reduce uncertainty by conforming to an accepted practice style (Gerrity et al., 1990).

In a FFS test-ordering paradigm, reducing a physician's uncertainty would mean ordering more laboratory and radiology tests. Many medical malpractice lawsuits (about 35.2% of payouts) relate to diagnostic errors (NASEM, 2016). A leading cause of such errors is failure by a medical professional to order a proper set of medical tests. If needed medical tests are not performed, a patient's health condition can be overlooked, which possibly leads to serious patient harm (Gandhi et al., 2006; Berlin, 2002). Thus, physicians tend to take proactive defensive attitudes and make efforts to avoid such errors by ordering a permissive, rather than restrictive, body of medical tests (Zhi et al., 2013). The more physicians in a hospital conform to this practice style, the higher overuse variation, or equivalently the lower underuse variation, in test-ordering practice will be. In payment reform model paradigms such as bundled payments, the care bundling contract will similarly standardize the recommended test panel. Physicians will follow evidence-based approaches to come to a diagnosis. Physician agreement about sufficient panels in medical tests should lead to less underuse/overuse variation in test-ordering practice and thus a proper amount of test-ordering cost.

While many experts highlight the tendency to overuse tests, underuse of test-ordering is also pervasive. Zhi et al. (2013) find not only a 30 percent overall rate of test overuse but also a similar rate of test underuse from a large-scale 15-year meta-analysis of laboratory testing practices. In this sense, high underuse variation in test-ordering practice implies that a hospital is more likely to fail to order proper tests for some patients. In a contract where standard evidence-based diagnosis processes are not followed, some physicians can more frequently underuse tests, while others still follow the tendency toward overuse, exhibiting practice variations. We conjecture that higher underuse variation in test-ordering practice can result in lower average test-ordering cost, yet worse system-wide operational performance (i.e., higher care-delivery cost), which we discuss next.

After a set of medical tests are given to a patient, no common strategy is known to reduce uncertainty like doing more. Instead, only a situational strategy exists: Do as your direct colleagues do (Eddy, 1984). We thus focus on underuse variation in test-ordering practice and its possible effect on subsequent care-delivery stages. Prior literature provides evidence that a small change in one process can cause enormous changes elsewhere. The bullwhip effect in supply chain (Lee et al., 1997), the snowball effect in communication networks (Krackhardt and Porter, 1986), and the butterfly effect in a complex system (Lorenz, 2000) are some examples. Hospital services also comprise a sequence of interrelated events (e.g., medical test, surgical procedure, recovery, etc.). When medical malpractice such as diagnosis error leads to incorrect treatment, delayed treatment, or no treatment, a patient's condition can be made even worse. For example, a patient with a heart attack that is misdiagnosed as psychiatric anxiety might only receive appropriate treatments after the symptoms aggravate overtime, and thus will require extra resources compared to a patient correctly diagnosed immediately. In this sense, we posit that higher underuse variation in test-ordering practice can lead to higher care-delivery cost:[5]

Hypothesis 2. *For a specific care episode, higher underuse variation in test-ordering practice (i.e., laboratory and radiology tests) relates to higher average care-delivery cost.*

[5] care-delivery cost = total cost - test-ordering cost

2.2.4.3 *Moderating Effects of Process Quality and Experiential Quality*

The practice style hypothesis (Wennberg and Gittelsohn, 1973) and the enthusiasm hypothesis (Chassin, 1993) are complementary theoretical foundations suggested to explain medical practice variation. Both theories argue that physicians follow their own preference for certain procedures, and thus they are known as a *preference-based approach*. The variations are suggested to stem from physicians' practical experience and educational history (Westert and Groenewegen, 1999).

The *constraint-driven approach*, also known as depend-on-conditions approach, is an alternative to the preference-based approach (Lindenberg, 1990). In this alternative theory, the differences in social context define opportunities and constraints, dominantly determining preferences or practice styles, rather than the differences in tastes or pure preferences. The social context includes incentives for certain options or restriction on behavioral choices. Assuming physicians are goal-oriented individuals who optimally decide their practice style to realize personal goals (e.g., financial benefit), Westert and Groenewegen (1999) relate social conditions, such as institutional influences and incentives, to the options that physicians choose. Preferences can be understood as instrumental goals pursued to achieve more general goals, within a certain combination of opportunities and constraints. As such, behavior can be predicted from the social conditions that affect the realization of instrumental goals.

Similar to the reasoning under the constraint-driven approach above, the process quality and experiential quality measures, which hospitals are required by CMS to self-report, can be seen as exogenous institutional influences. Hence, the levels of these two quality measures capture the eagerness of physicians and other stakeholders in a hospital to achieve their financial benefit while also complying with the government goal. Accordingly, each of the two quality measures may influence the options physician choose when providing services to patients under given social conditions, affecting the relationships described in Hypothesis 1 and Hypothesis 2. Hence, pursuing process quality and experiential quality may affect a hospital's operational performance (CMS, 2014b).

From Total Quality Management (TQM) literature, firms must coordinate behavioral, tacit, and intangible resources to improve processes (Powell, 1995). Decision makers may share tacit as well as explicit knowledge to maintain or improve their process performance (Benner and Tushman, 2003). The CMS process quality measure, which comprises broadly recognized clinical standards of care, entails using repetitive explicit and tacit practices to achieve process improvements. Such procedure standardization can reinforce the pattern recognition that humans do well, leading to lower errors (Leape, 1994). Thus, a hospital that has high scores on process quality is equipped with a more coordinated course of action (Chandrasekaran et al., 2012), which may weaken the relationship between practice variation and resource usage described in Hypothesis 1. We can similarly apply this argument to the relationship described in Hypothesis 2. That is, hospitals with both higher underuse variation in test-ordering practice and process quality would have reduced cases of unnecessary excessive care, leading to relatively lower care-delivery cost, compared to the hospitals with lower process quality.

Hypothesis 3a. *For hospitals with higher process quality (PQ), the positive relationships between overall practice variation and resource usage (i.e., H1) and between underuse variation in test-ordering practice and care-delivery cost (i.e., H2) become weaker.*

Meanwhile, we identify distinct performance aspects of experiential quality. Unlike process quality, which deals with attributes controllable by stakeholders inside a hospital, experiential quality is less controllable due to the patient engagement. Relying on organizational learning theory (Benner and Tushman, 2003; March, 1991), process quality is referred to as exploitation that reduces variation, whereas experiential quality is compared to patient-specific exploration that induces variation (Chandrasekaran et al., 2012). Higher experiential quality reflects managerial and caregiver efforts within a hospital to provide more responsive treatment for every patient. Hence, hospital managers that pay attention to provide better patient experience may be even more likely to devote time/effort/resources in their patient treatment services. Thus, we posit that the higher level of responsive care may amplify the positive link between practice variation and resource usage described in Hypothesis 1. We again similarly apply this argument to the relationship described in Hypothesis 2. Hospitals with both higher levels of underuse variation in test-ordering practice and experiential quality will have more cases of resource-intensive care, reflecting high responsiveness and a tendency to accommodate patient requests, leading to higher subsequent care-delivery costs.

Hypothesis 3b. *For hospitals with higher experiential quality (EQ), the positive relationships between overall practice variation and resource usage (i.e., H1) and between underuse risk in test-ordering practice and care-delivery cost (i.e., H2) become stronger.*

2.3 Data, Variables, and Model Development

We next describe data sources and discuss how we construct variables. We then develop econometric models to examine our hypotheses.

2.3.1 Description of Data Set

We use a comprehensive data set built by merging four databases. First, we use *State Inpatient Discharges* data from the *Healthcare Cost and Utilization Project (HCUP)* from New York (NY) and Florida (FL) states to identify patient-level information. NY and FL states are the second- and fourth-largest healthcare markets in the U.S., respectively, in terms of healthcare expenditures by state of residence (CMS, 2011).[6] Both states provide information to track readmitted patients. We use six years of data, from 2008 to 2013, the period for which the two quality metrics are fully available. Each yearly HCUP data file contains the domain of the inpatient discharge record, such as patient demographics, comorbidities, diagnoses, procedures, LOS, physician identifiers, payer, and claim charge information. We extract data records that relate to Medicare patients with conditions AMI, HF, and PN, since the data for the two quality measures are Medicare claims. Using the HCUP data, we construct measures of practice variation (i.e., our main independent variable), LOS, and cost components (i.e., dependent variables).

Second, we use *CMS Timely and Effective Care* data maintained by CMS to obtain process quality metrics. The measures apply to Medicare patients. From among the small list of conditions having process quality measures, we focus on AMI, HF, and PN. We construct a hospital-level composite score of process quality that we discuss later. To incentivize hospitals to participate in

[6] 1st: California $230,090; 2nd: New York $162,845; 3rd: Texas $146,735; 4th: Florida $132,463; 5th: Pennsylvania $97,414; $\cdots$; Overall in the U.S.: $2,089,862 (in 2009, in millions). Thus, New York and Florida states account for about 14.13% of total health spending in the U.S.

data collection processes, HHS withholds 0.4% of Medicare fees from the hospitals that choose not to participate in the program (Jha, 2006). Thus, we expect these data to be reasonably comprehensive. Third, we use *CMS HCAHPS* survey data to construct experiential quality measures. Fourth, we merge these databases with the *Historical Inpatient Impact File for Acute Inpatient Prospective Payment System*, which is maintained by CMS on an annual basis. We obtain the hospitals' structural characteristics including bed size, CMI, ownership type, corporate goals, location, and teaching intensity, each of which are used as control variables.

In our analysis, we use the merged dataset for 387 healthcare providers from NY and FL. Hospital structural characteristics are summarized in Table 2.1.[7] Figure 2.3 shows the process of reducing patient-level HCUP SID data to the dataset pertaining to the 387 hospitals. Our final 1,094,111 patient-level records include patients with three medical conditions (i.e., AMI, HF, and PN) who are Medicare beneficiaries and are not within the top or bottom 1% outliers for each state, year, and MS-DRG code, in terms of total medical charges.[8] Since process quality and experiential quality measures are not reported for medical conditions having less than 10 patients in a year within a hospital, the actual number of providers in our analyses is slightly less than 387 per year.

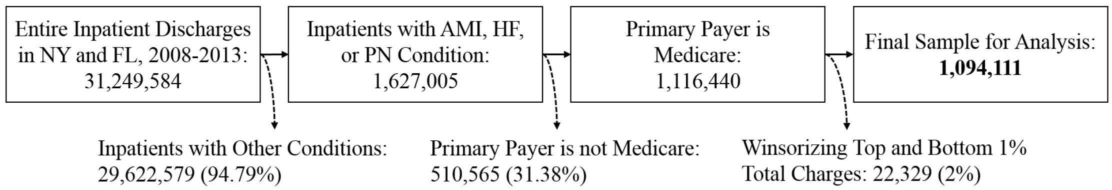
Table 2.1: Hospital Structural Characteristics

	Frequency	Percentage		Frequency	Percentage
Teaching Status			*State*		
Non-teaching	217	56.07	FL	188	48.58
Teaching	170	43.93	NY	199	51.42
Ownership and Corporate Goals			*Geographic Location*		
Governmental	56	14.47	Rural	52	13.44
For-profit	83	21.45	Urban	307	79.33
Non-profit	248	64.08	Missing	28	7.24

[7] *Note*: A hospital identifier called *DSHOSPID* in the HCUP SID data is used to match hospitals in Florida state (using *CMS Provider Number*) and New York state (using *NY SPARCS PFI (New York Statewide Planning and Research Cooperative System Permanent Facility Identifier)*).

[8] As a robustness check, we also considered top and bottom 3% outliers for each state, year, and MS-DRG code, resulting in consistent results.

Figure 2.3: Process of the Patient Sample Selection



2.3.2 Variables

The variables used in this study are summarized in Appendix A.2.

2.3.2.1 Dependent Variable: Hospital Operational Performance

We operationalize the dependent variables (i.e., *Total LOS* and *Total Cost*) by tracking patients' initial visits and revisits to hospitals within 30 days. *Total LOS* and *Total Cost* measures are often used to represent the level of resource usage. The total LOS (total cost) comprises (i) initial LOS (initial cost) during a patient's initial admission and treatment period and (ii) readmission LOS (readmission cost) of any unplanned readmissions within a 30-day post-discharge period.[9] Via this approach, we obtain more accurate estimates of resource usage during an entire care episode instead of for only an individual discharge (Andritsos and Tang, 2014a). The 30-day post-discharge window is commonly used in practice since hospitals consider the 30-day window as clinically meaningful and a long enough period for collaborations with care communities to reduce readmissions (Drye et al., 2012).

Cost measures are estimated from patient-level charge data in HCUP SID by following the approach suggested in Chen et al. (2010) and also promoted by the Agency for Healthcare Research and Quality. We first apply a national consumer price index for hospital services to the charge data to convert them into 2013 U.S. dollars. We then multiply the inflation-adjusted charges by the CMS annually announced cost-to-charge ratio to estimate each hospital's inpatient operating costs.

Further, we construct a risk-adjusted hospital-level mean value of each dependent variable. In other words, we adjust for the warranted variation, to compare hospitals in a fair manner (Andrit-

[9]See Appendix A.3 for different cases of patient revisit from HCUP SID Data.

sos and Tang, 2014b; CMS, 2015b). This process removes the influence of patient characteristics that possibly affect LOS and care-delivery cost for reasons not related to practice variation or care quality of a focal hospital. Specifically, we control for patient demographics (i.e., age, gender, and race), admission-type indicators (i.e., emergency, urgent, and elective), and 29 comorbidity indicators, consistent with recommendations from CMS, National Quality Forum, and the American Heart Association (Horwitz et al., 2011). This process requires estimation of the following model:

$$y_{ijkt} = \alpha + \beta \mathbf{P_k} + \gamma \mathbf{A_k} + \delta \mathbf{C_k} + \epsilon_{ijkt}, \tag{2.1}$$

where y_{ijkt} is either $ln(TotalLOS)_{ijkt}$ or $ln(TotalCost)_{ijkt}$ for a patient k with condition j who visited hospital i in year t. P_k, A_k, and C_k are vectors of patient k's demographic factors, admission-type indicators, and comorbidity indicators, respectively.[10] Then, we calculate the predicted dependent variable, $\hat{y}_{ikt}$, for each i, k, and t using the estimated model of Equation (2.1). We then obtain hospital i's mean predicted dependent variable, i.e., $\hat{y}_{it}$, by taking the average across patients in a focal hospital. Similarly, we also calculate hospital i's mean observed dependent variable, i.e., y^o_{it}, by averaging the actual dependent variables of patients in a hospital. Lastly, we obtain the hospital-level risk-adjusted dependent variable as $y_{it} = \frac{y^o_{it}}{\hat{y}_{it}} \cdot \bar{y}$ where $\bar{y}$ is the mean of the actual dependent variable across all patients and years used to estimate Equation (2.1).

2.3.2.2 Independent Variable: Practice Variation

To measure the degree of practice variation in a hospital, we construct a metric called weighted average coefficient of variation (WACV) for each year and care episode (e.g., AMI, HF, and PN) treatment within a hospital.[11] Our main purpose of this measure is to track practice variation for a clinically coherent set of patients that enables us to relate a hospital's observed outcomes for each condition to the resource demands and associated costs experienced by the hospital.

[10] As previous literature pointed out significant geographic variation in cost and charges (e.g., Fisher et al., 2003; Miller et al., 2011), we estimate a separate model of Equation (2.1) for each condition and for each state.

[11] We can also derive this measure for each year, hospital, and condition (i.e., AMI, HF, and PN), and their histograms are shown in Figure A.4. We use this condition-level practice variation measure in the analysis by condition discussed in Appendix A.10.

When any Medicare beneficiary discharges from a hospital, a single MS-DRG[12] code is assigned to that patient, which is determined based on multi-dimensional information including principal diagnosis, secondary diagnoses, surgical procedures, age, gender, and discharge status of the patient. For each medical condition, in general, there are three MS-DRG codes to differentiate patients in terms of illness severity, prognosis, and treatment difficulty (CMS, 2016b). For example, an AMI patient can be assigned to MS-DRG 280 (i.e., with Major Comorbidities, wMCC), 281 (i.e., with Comorbidities, wCC), or 282 (i.e., without Comorbidities, woCC). As such, the patients assigned to the same MS-DRG code are expected to receive similar therapeutic and bed services used in the management of a particular disease, and hence become a homogeneous group in terms of payments. Therefore, if the amounts of total charges vary substantially for the patients within such a group, then we conclude that the degree of practice variation of the group is high.

To generate WACV, we calculate the coefficient of variation (CV) for patients within a same MS-DRG code, and then obtain a weighted-average value at the hospital-level, with the weight as the number of patients in each code (see Figure A.1 in Appendix A.4). Let t denote year (i.e., $t \in T = \{2008, \cdots, 2013\}$), i a hospital in the integrated dataset, j an element of the set of conditions (i.e., $j \in J = \{AMI, HF, PN\}$), and s a MS-DRG code that reflects illness severity (i.e., $s \in S_j = \{wMCC, wCC, woCC\}$). The CV is the ratio of the sample standard deviation sd_{ijst} to the sample mean $\bar{x}_{ijst}$ of total medical charges for each i, j, s, and t (i.e., $CV_{ijst} = sd_{ijst}/\bar{x}_{ijst}$ where $sd_{ijst} = \frac{1}{n_{ijst}-1} \sum_{k}^{n_{ijst}} (x_{ijst}^{k} - \bar{x}_{ijst})^2$ where k indicates patient.).[13]

The CV is useful since it is independent of the unit in which the measurement is taken, whereas the usual standard deviation measure must always be understood in the context of the mean of the data. For this reason, to compare multiple data sets (e.g., hospitals) having far different means or different units, CV is recommended. In the context of healthcare spending, various adjustment factors, such as wage index and CMI, are typically included in empirical models to rule out possible differences across regions and locations. When we derive practice variation via CV, which is a

[12]MS-DRG: Medicare Severity-Diagnosis Related Group.

[13]Depending on the "target value" that we discuss in Appendix A.9, $\bar{x}_{ijst}$ can be defined at either hospital-, county-, or state/CBSA-level.

dimensionless number, the value itself is robust for comparisons between and within hospitals, even if we do not perform the wage index or CMI adjustments. WACV for hospital i in year t is computed as:

$$WACV_{it} = \frac{\sum_{j \in J} \sum_{s \in S_j} N_{ijst} \cdot CV_{ijst}}{\sum_{j \in J} \sum_{s \in S_j} N_{ijst}}, \tag{2.2}$$

where N_{ijst} is the number of patients and CV_{tijs} is defined as $sd_{tijs}/\bar{x}_{tijs}$ for each $i,\ j,\ s$, and t.

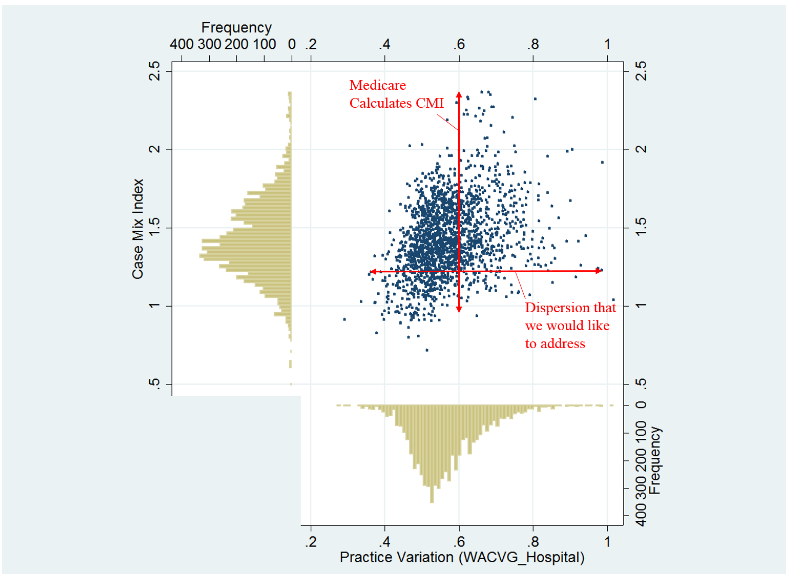
The concept of a weighted average of CV across illness severities is adopted to capture variations after controlling for the medical condition, illness severity, and treatment complexity as much as possible.[14] Table A.4 in Appendix A.5 contains the number of patient cases, summary statistics for age, gender, comorbidities, diagnoses, procedures, LOS, and total charges across MS-DRG codes for each condition. From this table, we observe that the total charges might depend on the level of illness severity and treatment complexity. That is, the higher the severity level and complexity, the higher the total charges, because hospitals provide more care services to those patients. Hence, if we derive CV for each entire health condition (e.g., one group made up of all patients with heart failure, regardless of illness severity), then the magnitude of CV can be highly affected by the distribution of patient cases. In essence, both warranted and unwarranted variation in claim charges would be captured by such a metric, which makes it difficult to interpret the magnitude of CV. Hence, we use WACV instead, based on MS-DRG codes as the boundary of patient groups. We claim that higher WACV values indicate higher practice variation. To further rule out other sources of warranted variation (e.g., age, gender, and comorbidities as listed in Table A.4, and admission type and race as listed in Table A.5 in Appendix A.5) that are not directly related to the "level of practice," we go through a risk-adjustment process for the charge measures, as previously discussed in Section 2.3.2.1.

Compared to CMI, the WACV metric contributes by enabling appropriate analysis of within-hospital procedural variability after risk-adjusting for patient diversity. Indeed, as Figure 2.4

[14]Constructing a weighted-average CV across conditions enables us to flexibly change the level of analysis. For example, WACV for hospital i and condition j in year t can be similarly computed as $WACV_{ijt} = (\sum_{s \in S_j} N_{ijst} \cdot CV_{ijst})/(\sum_{s \in S_j} N_{ijst})$, of which the estimation results are included in Appendix A.10.

shows, the level of practice variation can vary significantly (e.g., [0.36, 1.00]) even when hospitals have the same level of CMI (e.g., 1.25). The histogram of practice variation in Figure 2.4 is roughly bell-shaped. We note that the hospitals located toward the left side of the histogram (e.g., less than 0.5) have relatively lower practice variation compared to those located toward the right side.

Figure 2.4: Practice Variation vs. Case Mix Index



We also carefully rule out possible limitations of a measure based on CV. If the sample size for each MS-DRG code is too small, then the reliability of WACV may be problematic (Kelley, 2007). Therefore, we include hospital-level observations only if the number of patient discharges per condition per fiscal year is greater than or equal to 25. This approach is in line with CMS recommendations (Nair et al., 2013). For robustness purposes, we also test the same regression models by including all of the observations, which consistently shows the same results. We also calculated WACV using an alternative boundary of patient group (i.e., principal diagnosis within each medical condition, instead of MS-DRG code), and again obtain consistent results. We discuss in detail these robustness checks in Section 2.4.4.

Thus far, we have discussed the construction of the overall practice variation variable used to examine Hypothesis 1. For Hypothesis 2, we need separate measures for underuse and overuse variation in test-ordering practice. We adopt a concept from Finance called target semi-deviation (TSD) that is computed in a similar way to a standard deviation, but only on a single side of a target mean (Estrada, 2007; Rohatgi, 2011). Specifically, we take the sum of squares of differences from the target a (e.g., mean), divide by the number of observations, and take a square-root. In computing the lower TSD (for *underuse variation in test-ordering practice*), we use 0 in place of deviations above the target; and in computing the upper TSD (for *overuse variation in test-ordering practice*), we use 0 in place of deviations below the mean. Equations make this more clear:

$$TSD_{lower}(X,a) = (\mathbb{E}[(X-a)^2 \cdot \mathbf{1}_{\{X \leq a\}}])^{1/2}, \tag{2.3}$$

where a is a target (e.g., $\mathbb{E}[X]$) and $\mathbf{1}_{\{X \leq a\}}$ is an indicator function, i.e., $\mathbf{1}_{\{X \leq a\}} = \begin{cases} 1 & \text{if } X \leq a; \\ 0 & \text{else.} \end{cases}$ Notice that values on the "opposite" side of the target are not simply omitted; rather, they are replaced by zeros, so the denominator of the TSD is the same as the denominator of the standard deviation. Hence, $Var(X,t) = TSD^2_{lower}(X,t) + TSD^2_{upper}(X,t)$ is satisfied. We replace TSD_{lower} (TSD_{lower}, respectively) with the sample standard deviation in the formula of $WACV_{\text{overall}}$ in Equation (2.2) to calculate $WACV_{\text{underuse}}$ ($WACV_{\text{overuse}}$, respectively). Figure 2.5 clarifies the boundary of practice variation measures in our study. Using data from patients with pneumonia in year 2011, Figure 2.6 illustrates how we classify patients into either the underuse case or the overuse case. If a patient exhibits a test-ordering charge below (above, resp.) the within-hospital average (i.e., the lines in Figure 2.6), the patient falls into the underuse case (the overuse case, resp.). As robustness checks, we also use county-average and state/CBSA-average as different targets, which show largely consistent results (see Appendix A.9).

2.3.2.3 Moderators: Process Quality and Experiential Quality

We follow an approach suggested by CMS to derive the composite process quality (PQ) score from individual measures via a weighted average approach. The size of the eligible patient pop-

Figure 2.5: Boundary of Practice Variation Measures

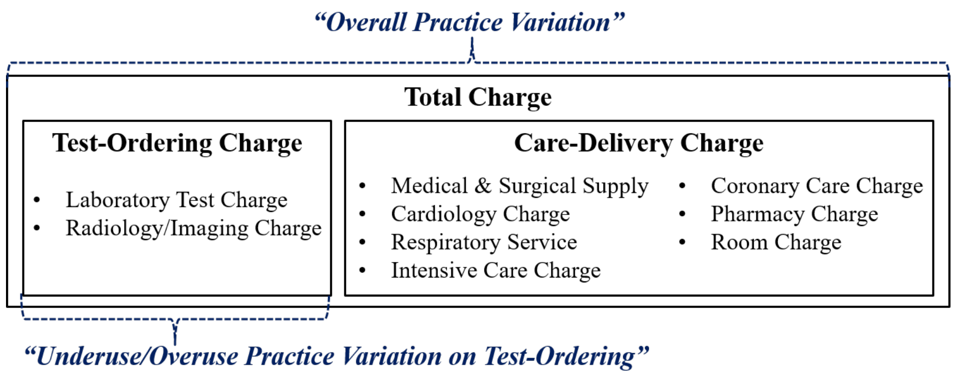
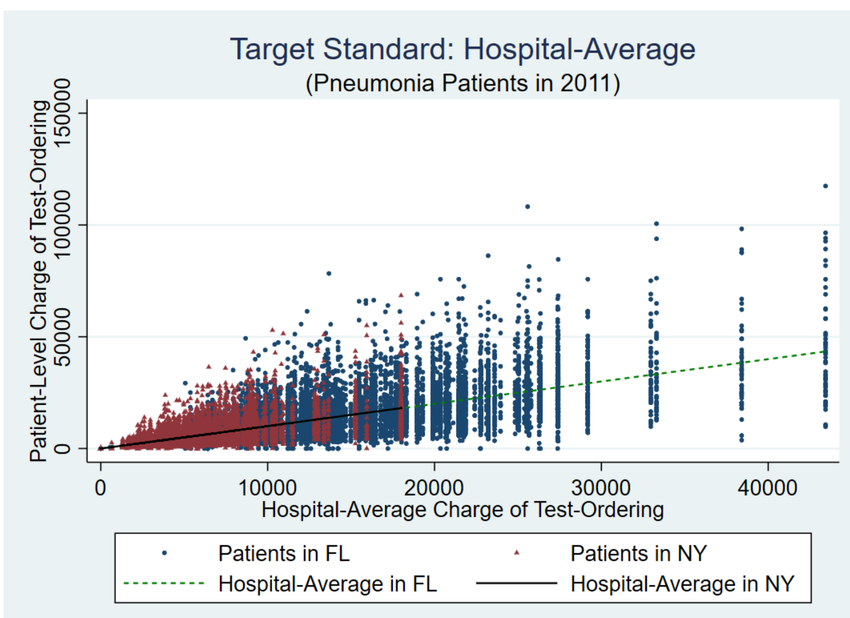


Figure 2.6: Variation in Test-Ordering Practice, Target: Hospital-Average



ulation becomes a weight for each measure. Prior literature also uses this approach (e.g., Chandrasekaran et al., 2012). Specifically, we use 12 quality measure items for AMI, 4 items for HF, and 7 items for PN to generate composite PQ for each hospital or for each condition, where the list of items slightly varies from year to year. The descriptions and summary statistics for the measure items are provided in Appendix A.6. We interpret the resulting value as a compliance score that reflects the extent to which medical guidelines are followed. Let m denote a process quality measure item and M_j denote the set of measures for condition j. The number of patients differs across measures since there may be some reasons that some patients do not need to receive a certain treatment. We derive the weighted average process quality score per hospital for each year as follows:[15]

$$PQ^o_{it} = \frac{\sum_{j\in J}\sum_{m\in M_j} N_{ijmt}\cdot q_{ijmt}}{\sum_{j\in J}\sum_{m\in M_j} N_{ijmt}}, \tag{2.4}$$

where q_{ijmt} is the associated process quality score and N_{ijmt} is the number of patients for each i, j, m, and t.[16] The distribution of each q_{ijmt} in the CMS database and the resulting PQ^o_{it} measures are left-skewed. Hence, we perform a logit transformation of the odds ratio of PQ^o_{it} to make this highly skewed distribution less skewed (Cohen et al., 2003):

$$PQ_{it} = ln\left(\frac{PQ^o_{it}}{1-PQ^o_{it}}\right). \tag{2.5}$$

In a similar vein, we derive the composite experiential quality (EQ) score (e.g., Nair et al., 2013). The experiential quality score is based on six items in the HCAHPS survey (see Appendix A.6). Within the survey, the responses for the first five items (i.e., Comp1-Comp5) are presented as "Never/Sometimes," "Usually," or "Always," and for the sixth item (i.e., Comp6) the response is reported as "Yes" or "No." To handle this difference in the data structure for the first five items,

[15]In calculation of a composite process quality measure, we drop AMI-4, HF-1, HF-4, and PN-4, which are more related to advice/counseling or instructions rather than clinical process (Andritsos and Tang, 2014b). However, our estimation results remain consistent even when including these measures.

[16]We can also construct a condition-level measure, which is used in robustness checks, for each hospital and year: $PQ^o_{ijt} = \frac{\sum_{m\in M_j} N_{ijmt}\cdot q_{ijmt}}{\sum_{m\in M_j} N_{ijmt}}$.

we use the percentage of patients who answered "Always" as the measure of each individual item (Senot et al., 2015). For the sixth item, we designate the percentage of patients who answered the question with "Yes" as a final score. Lastly, we compute the hospital-level overall score as the average of the percentage scores for the six items ($i.e., EQ^o_{it}$). Similar to the process quality score, the experiential quality score for each hospital is then calculated by taking the logit transformation of the percentage. Thus, EQ_{it} for a hospital i and year t is given by:

$$EQ_{it} = ln\left(\frac{EQ^o_{it}}{1 - EQ^o_{it}}\right). \tag{2.6}$$

To obtain easy-to-interpret results, we mean-center the practice variation measure and quality measures before computing their interaction terms (Hamilton, 2012).

2.3.2.4 Controls

We control for several hospital factors that are related to potential sources of heterogeneity in performance across hospitals. Hospital size is measured according to its total number of beds. We take the logarithm to account for heavy tails in this distribution ($BedSize$). CMI captures the average DRG weight for different DRGs per hospital (Ding, 2014). CMS derives CMI by calculating the ratio between the total DRG weights associated with Medicare discharges and the total discharges. The $TeachingIntensity$ of a hospital is defined as the residents-to-bed ratio (Theokary and Ren, 2011). $WageIndex$ reflects the relative hospital wage level in the geographic area of the hospital compared to the national average hospital wage level (Shwartz et al., 2011). We also control for the CMS operating outlier adjustment factor ($OutlierAdjustment$), which reflects the extent of uncommonly costly patients treated by the focal hospital, and the CMS operating disproportionate share hospital payment adjustment factor ($OPDSHAdjustment$), where the disproportionate share hospitals serve a notably disproportionate number of low-income patients and receive payments from CMS to cover the costs of providing care to uninsured patients (Senot et al., 2015).

The control variables so far are all time-varying, which is appropriate for the dynamic panel model that we discuss later. We also use time-invariant controls for patient-level analysis in robustness checks and static OLS estimates in post-hoc analyses (see Appendices A.13 and A.14). Each hospital is classified into three types: government-sponsored, private non-profit, and private for-profit. Setting for-profit hospitals as the base group, we use two binary variables, one for private non-profit hospitals ($Nonprofit$) and another for government-sponsored hospitals ($Governmental$). Hospital location is classified as either urban (1) or rural (0) in variable $Urban$. We also include year dummies to control for unobservable factors that cause population change in hospital operational performance.

2.3.3 Methodologies and Econometric Models

We discuss the dynamic relationship between practice variation, quality, and operational performance, and develop econometric models to examine our hypotheses.

2.3.3.1 A Dynamic Model of Operational Performance

Endogeneity is pervasive across many aspects of healthcare operations. The specific effect of endogeneity may arise from the dynamic relationship between current hospital operations and a hospital's history. We examine how practice variation as well as quality measures relate to operational performance. The level of the two quality measures is dynamically endogenous with respect to operational performance because manager talent can affect quality measures that are closely linked with financial incentives. We believe the relationship between practice variation and operational performance is similar.

As Section 2.2 implies, practice variation and the two quality measures are choice-type variables in a broad sense, arising through a process of bargaining between the decision makers inside a hospital (e.g., board members, physicians, administrative staffs). Although governmental policies might be the most critical driver, this process is also influenced by past performance, manager talent, and decision makers' beliefs about the benefits and cost of choosing reasonable clinical pathways for patients, leading to various levels of practice variation and quality across hospitals. Therefore, if practice variation and quality measures are dynamic, and hospital i (given its performance at time $t-1$ or earlier) chooses certain levels of practice variation and quality score $\mathbf{X_{it}}$ to achieve a particular level of expected operational performance at time t, then the dynamic model for practice variation and quality is:

$$\mathbf{X_{it}} = f(y_{it-1}, \cdots, y_{it-p}, \mathbf{Z_{it}}, \eta_i), \tag{2.7}$$

where $\mathbf{X_{it}}$ is a vector of endogenous predictors (i.e., practice variation, quality measures, and their interaction terms) that we call *practice-quality status*, $\mathbf{Z_{it}}$ is a vector of exogenous predictors (i.e., controls and time dummies), y_{i*} represents operational performance, and η_i is an unobserved hospital effect.

Equation (2.7) suggests that estimating the effect of *practice-quality status* on operational performance, conditional on hospital heterogeneity, requires estimating the following model:

$$y_{it} = \alpha + \sum_p \lambda_p \cdot y_{it-p} + \beta \cdot \mathbf{X_{it}} + \gamma \cdot \mathbf{Z_{it}} + \eta_i + \epsilon_{it} \tag{2.8}$$

where ϵ_{it} is an idiosyncratic error term and β is the coefficients of interest. To estimate Equation (2.8), we assume that current shocks are independent of historical realizations of performance and *practice-quality status*. In other words, past and current realizations of practice variation and quality scores are allowed to influence current performance. This assumption is not too strong and leaves open the possibility that hospitals may strategically choose their level of practice variation and target quality to affect current or future performance. If the level of *practice-quality status* that we observe today is the one that trades off the anticipated benefits and costs of doing so, then the unanticipated component of performance, many years in the future, will not be related to the *practice-quality status* that is realized today. Intuitively, we can write this in orthogonality form as $E(\epsilon_{it}|y_{it-s}, \mathbf{X_{it-s}}) = 0, \forall s > p$. If Equation (2.7) represents the true model for performance, that is, if we correctly identified every endogenous time-varying variable that affects performance, then ϵ_{it} is an expectational error and the orthogonality assumption is valid (Hansen and Singleton, 1982). Equation (2.8) is simply a reduced-form model and thus the reduced-form error, ϵ_{it}, is at best a proxy for the pure expectational error (Wintoki et al., 2012).

However, there are several challenges in empirically estimating the fixed-effect model (Pang et al., 2016). As the hospital operational performance is likely to be affected by hospital-specific unobserved heterogeneity (η_i) that may be correlated with explanatory variables, hospital time-invariant fixed-effects need to be accounted for. However, fixed-effects estimation does not completely control for the correlation between η_i and the lagged dependent variable[17] (Roodman, 2006). Hence, we estimate a dynamic panel data model via System GMM estimation.

2.3.3.2 Estimation Strategy: Dynamic Panel System GMM

Under the assumption that unobserved heterogeneity is time-invariant, we obtain consistent and unbiased estimates of the relationship between *practice-quality status* and operational performance via a dynamic panel System GMM estimator (Arellano and Bover, 1995; Blundell and Bond, 1998). This estimator exploits the dynamic relationships inherent in our independent variables. The dynamic modeling approach has been widely used in areas such as economics and finance, where the structure of the problem contains a dynamic relationship between independent and dependent variables (e.g., Bond and Meghir, 1994; Blundell and Bond, 1998). Recently, operations management, information systems, and marketing fields also began examining relevant problems using the dynamic panel GMM approach (e.g., Narayan and Kadiyali, 2015; Senot et al., 2015; Bhargava and Mishra, 2014; Rego et al., 2013).

The estimation comprises two steps. First, we write the first-differenced form of Equation (2.8):

$$\Delta y_{it} = \alpha + \lambda_p \cdot \sum_p \Delta y_{i,t-p} + \beta \cdot \Delta \mathbf{X_{it}} + \gamma \cdot \Delta \mathbf{Z_{it}} + \Delta \epsilon_{it}, \text{ where } p > 0. \tag{2.9}$$

The first-difference effectively removes any bias that may arise from unobserved time-invariant heterogeneity. After first-differencing, we estimate Equation (2.9) via GMM using lagged values of performance, practice variation, quality scores, and other hospital-specific variables as instruments

[17]No matter how we try to get close to the “true" model (e.g., by adding controls that determine *practice-quality status*), we cannot completely rule out the possibility that we have omitted an endogenous time-varying variable that has an empirically significant effect on both operational performance and *practice-quality status*.

for current changes in these variables. An essential aspect of the dynamic panel estimator is its use of hospital-level history as instrument variables for our explanatory variables. Thus, in estimating Equation (2.8) or the first-difference as in Equation (2.9), our instruments will be drawn from the set of lagged dependent or independent variables (i.e., y_{it-k}, $\mathbf{X_{it-k}}$, $\mathbf{Z_{it-k}}$, where $k > p$.). For these instruments to be valid, they should satisfy two criteria. First, they must provide a source of variation for current *practice-quality status* (i.e., $\mathbf{X_t} = f(y_{t-k}, \mathbf{X_{t-k}}, \mathbf{Z_{t-k}})$). We show later that practice variation and quality scores are correlated to lagged operational performance and lagged values of other control variables (see Table A.7 in Appendix A.7).

Second, the historical values of practice variation and quality scores must explain an exogenous source of variation for current *practice-quality status*. As such, the lagged variables should not be correlated with the error term in Equation (2.8). Any information from the prior p periods is reflected in the current expected performance. Thus, p lags of performance are enough to address the impact of the hospital's past on the present. The hospital's history beyond period $t-p$ should be exogenous with respect to any shocks to performance in the current and future periods. Under the exogeneity assumption, then the following orthogonality condition is valid (Wintoki et al., 2012):

$$E(\mathbf{X_{it-s}}\epsilon_{it}) = E(\mathbf{Z_{it-s}}\epsilon_{it}) = E(y_{it-s}\epsilon_{it}) = 0, \ \ \forall s > p. \tag{2.10}$$

The number of lags included for each dependent variable in our analysis is revealed to be one according to how many lags are statistically significant in the corresponding regression.[18]

We then estimate the level and difference equations simultaneously, as Arellano and Bover (1995) and Blundell and Bond (1998) show that the GMM estimator can be improved compared to solely estimating a first-difference model. We use the first-differenced variables as instruments

[18]We tested the same model with different number of lags for each dependent variable to determine the number of lags to be included in our analysis.

for the level equations in a system of equations as below (Roodman, 2006):

$$\begin{bmatrix} y_{it} \\ \Delta y_{it} \end{bmatrix} = \alpha + \lambda \begin{bmatrix} \sum_p y_{it-p} \\ \sum_p \Delta y_{it-p} \end{bmatrix} + \beta \begin{bmatrix} \mathbf{X_{it}} \\ \Delta \mathbf{X_{it}} \end{bmatrix} + \gamma \begin{bmatrix} \mathbf{Z_{it}} \\ \Delta \mathbf{Z_{it}} \end{bmatrix} + \epsilon_{it}, \tag{2.11}$$

where y_{it} = *ln(Average Total LOS)* or *ln(Average Total Cost)* for a hospital i in year t, $\mathbf{X_{it}} = [\text{WACV}, \ \text{PQ}, \ \text{EQ}, \ \text{WACV} * \text{PQ}, \ \text{WACV} * \text{EQ}]$, and $\mathbf{Z_{it}}$ = a set of time-varying controls. Variable y_{it} and Δy_{it} denote the level and year-to-year change (from $t-1$ to t) in operational performance in hospital i. The level and change in practice variation, quality scores, and their interactions are captured by $\mathbf{X_{it}}$ and $\Delta \mathbf{X_{it}}$.

Note, however, that the level equations still have unobserved heterogeneity, η_i. As in Wintoki et al. (2012) and Kuhnen and Niessen (2012), we assume that the correlation between *practice-quality status* and control variables is constant over time. Relying on this assumption, we have another set of orthogonality conditions:

$$E[\Delta \mathbf{X_{it-s}}(\eta_i + \epsilon_{it})] = E[\Delta \mathbf{Z_{it-s}}(\eta_i + \epsilon_{it})] = E[\Delta y_{it-s}(\eta_i + \epsilon_{it})] = 0, \ \ \forall s > p. \tag{2.12}$$

We check the validity of instruments $\mathbf{Z_{it}}$ with serial correlation tests and the Hansen test of over-identification (Arellano and Bond, 1991) and show the test statistics in the results tables (i.e., Table 2.3 and 2.4). According to the serial correlation tests, the assumptions of our specifications are valid, that is, the residuals in first-difference ($AR(1)$) are significantly correlated, but there is no serial correlation in second-differences ($AR(2)$). In addition, the Hansen test with insignificant p-values in all specifications indicates that the null hypothesis that our instruments are valid is not rejected. Lastly, the difference-in-Hansen test tells us that the subset of instruments used in the level equations is exogenous for all model specifications.

2.4 Model Estimation Results

Table 2.2 reports summary statistics and correlations of key variables.

Table 2.2: Descriptive Statistics and Correlation Table of Key Variables

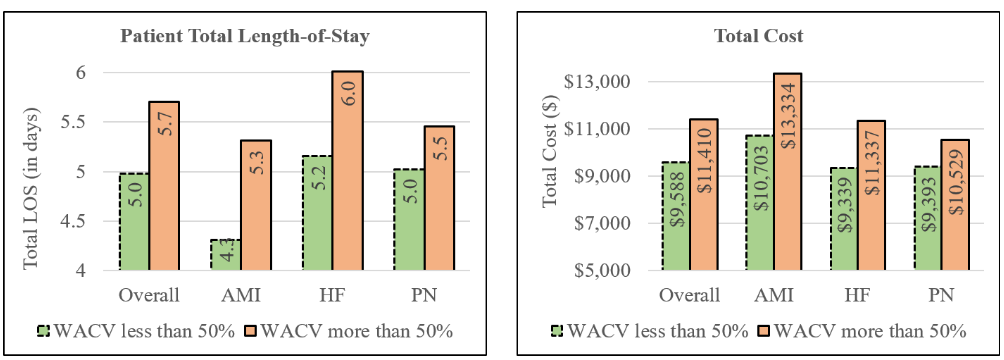
	Variable	n	Mean	S.D.	Min	Max	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
(1)	ln(Total LOS)	2045	1.694	0.219	-0.667	2.933	1														
(2)	ln(Total Cost)	2032	9.092	0.424	6.146	10.322	0.452	1													
(3)	ln(Cost of Ordering Test)	2025	7.74	0.545	2.32	9.232	0.024	0.639	1												
(4)	ln(Cost of Care Delivery)	2032	8.872	0.443	6.015	10.33	0.518	0.92	0.352	1											
(5)	WACVG	2004	0.566	0.093	0.272	1.017	0.426	0.205	-0.124	0.309	1										
(6)	WACVG_Underuse	2017	0.323	0.053	0	0.861	0.13	-0.093	-0.375	0.078	0.357	1									
(7)	WaCVG_Overuse	2017	0.495	0.127	0	2.74	0.164	-0.104	-0.413	0.086	0.4	0.889	1								
(8)	ProcQual	2356	-0.017	1.239	-6.099	5.223	0.053	0.078	0.193	-0.022	0.116	-0.002	0.004	1							
(9)	ExpeQual	2265	0	0.282	-0.86	1.653	-0.346	-0.143	-0.026	-0.124	-0.263	-0.097	-0.124	-0.092	1						
(10)	ln(Bed Size)	2484	5.148	0.96	1.386	7.564	0.329	0.255	0.192	0.194	0.482	0.141	0.196	0.462	-0.391	1					
(11)	CMI	2329	1.42	0.246	0.488	2.366	0.113	0.193	0.164	0.16	0.317	0.131	0.126	0.433	-0.105	0.681	1				
(12)	Teacing Intensity	2329	0.108	0.208	0	0.961	0.304	0.222	-0.19	0.345	0.569	0.29	0.327	0.028	-0.299	0.385	0.289	1			
(13)	Wage Index	2311	1.026	0.174	0.814	1.493	0.45	0.519	-0.006	0.639	0.423	0.195	0.223	0.019	-0.298	0.264	0.12	0.569	1		
(14)	Outlier Adj Factor	2329	0.04	0.048	0	0.658	0.099	0.272	0.18	0.258	0.316	0.066	0.056	0.12	0.022	0.373	0.408	0.272	0.112	1	
(15)	OPDSH Adj Factor	2329	0.141	0.173	0	0.86	0.222	0.049	-0.208	0.126	0.353	0.185	0.19	0	-0.324	0.283	-0.026	0.612	0.445	0.113	1

n: the number of hospital-year observations (2008-2013).
We use the hospital-average based ex-post type of practice variation measures (i.e., (5)-(7)) in our main analysis and conduct robustness check based on both county-average and state/CBSA-average and also for the ex-ante type of measures (See Appendix A.7).
Notes: $p < 0.05$ if $|r| > 0.048$. n is the number of hospital-level observations for six years (2008 to 2013).

2.4.1 Model-Free Evidence

Before we present the results of our proposed model estimated using dynamic panel system GMM, we provide model-free evidence of the practice variation effect on hospital operational performance. As part of our model-free analyses, we present plots of practice variation and patient total LOS/total cost. For brevity, we compare the average per patient total LOS and total cost across two scenarios: hospitals with (1) practice variation less than the median in a year, and (2) practice variation more than the median in a year. As seen in Figure 2.7, we find that hospitals with a relatively higher level of practice variation exhibit longer patient LOS and higher total cost. We further shed light on the practice variation effect on test-ordering cost and care-delivery cost. As in Figure 2.8, hospitals with relatively higher practice variation tend to spend less on test-ordering activities while spending more in care-delivery activities.

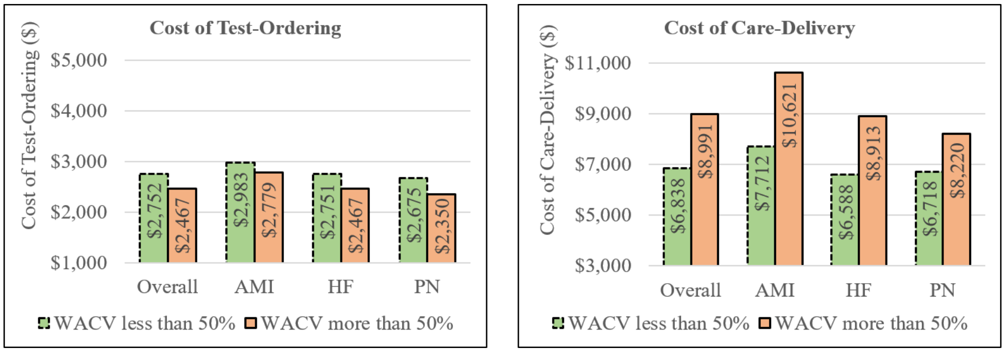
Figure 2.7: Model-Free Evidence of H1: Main Effect of Practice Variation on Risk-Adjusted Average Length-of-Stay (Left) and Total Cost (Right)



2.4.2 Impact of Practice Variation on Length-of-Stay and Total Cost per Capita

Table 2.3 summarizes the hospital-level estimates obtained for H1 together with H3a and H3b. With total LOS as a dependent variable, we first run a model with Process Quality and Experiential Quality (M1), and then include Practice Variation (M2). Lastly, we include the interaction terms (M3). We similarly examine the model with the same explanatory variables but with Total Cost

Figure 2.8: Model-Free Evidence of H2: Main Effect of Practice Variation on Risk-Adjusted Average Cost of Test-Ordering (Left) and Cost of Care-Delivery (Right)



per capita as a dependent variable (M4 to M6). All the instrument validity and identification tests reported in Table 2.3 support the use of dynamic panel system GMM to estimate the models.

In the relationship of interest in H1, the main-effect results (M3) in Table 2.3 show a significant positive association between practice variation (i.e., WACVG) and total LOS ($\beta = 0.635, p < 0.01$). This result indicates that patients staying at hospitals with greater practice variation tend to stay longer during their entire episode of care.

Consistent with previous literature (e.g., Nair et al., 2013), patients at hospitals with higher experiential quality tend to stay a shorter period ($\beta = -0.549, p < 0.01$). However, we do not observe any significant relationship between process quality and patient total LOS ($\beta = -0.073, p > 0.10$) although the direction is aligned with prior findings (e.g., Andritsos and Tang, 2014b).

H3a posits that the relationship between practice variation and total LOS/cost is weaker when process quality is high. Similarly, H3b suggests the relationship between practice variation and total LOS/cost is stronger when experiential quality is high. M3 in Table 2.3 shows the results on total LOS. The interaction between practice variation and process quality is not significant ($\beta = 0.130, p > 0.10$) with patient total LOS, indicating H3a is not supported. In contrast, the interaction between practice variation and experiential quality shows a significant positive association ($\beta = 0.920, p < 0.05$) with total LOS, providing support to H3b. Thus, the benefit of higher experiential quality in reducing patient total LOS can diminish once we explicitly consider

Table 2.3: Hospital-Level Results of System GMM (WACVG, Dep: Total LOS, Total Cost)

Dep Var	Total LOS			Total Cost		
	(M1)	(M2)	(M3)	(M4)	(M5)	(M6)
ProcQuality (PQ)	-0.001	-0.001	-0.073	0.004	0.004	-0.015
	(0.005)	(0.003)	(0.047)	(0.010)	(0.008)	(0.080)
ExpQuality (EQ)	-0.019	-0.050^{*}	-0.549^{***}	0.052	0.020	-0.134
	(0.050)	(0.028)	(0.205)	(0.134)	(0.071)	(0.171)
WACVG		0.553^{***}	0.635^{***}		0.318^{**}	0.233^{*}
		(0.060)	(0.070)		(0.127)	(0.122)
PQ*WACVG			0.130			0.027
			(0.084)			(0.141)
EQ*WACVG			0.920^{**}			0.291
			(0.369)			(0.338)
Teaching Intensity	0.014	-0.056^{**}	-0.054	0.021	-0.031	-0.010
	(0.023)	(0.025)	(0.038)	(0.063)	(0.048)	(0.055)
Bed Size	0.010	-0.002	-0.001	0.006	-0.002	-0.007
	(0.007)	(0.006)	(0.008)	(0.022)	(0.016)	(0.015)
Case Mix Index	0.010	-0.002	-0.017	-0.029	-0.028	-0.012
	(0.021)	(0.020)	(0.027)	(0.057)	(0.042)	(0.039)
Wage Index	0.064^{**}	0.046^{*}	0.113^{***}	0.450^{*}	0.352^{***}	0.327^{**}
	(0.031)	(0.027)	(0.034)	(0.232)	(0.120)	(0.139)
OPDSH Adj Factor	-0.011	-0.022	0.023	-0.248^{**}	-0.208^{***}	-0.175^{**}
	(0.028)	(0.028)	(0.035)	(0.106)	(0.074)	(0.076)
Outlier Adj Factor	-0.060	-0.154^{*}	-0.114	0.388	0.230	0.209
	(0.089)	(0.086)	(0.105)	(0.349)	(0.224)	(0.216)
Dep Var$_{(t-1)}$	0.779^{***}	0.705^{***}	0.551^{***}	0.673^{***}	0.721^{***}	0.739^{***}
	(0.065)	(0.049)	(0.064)	(0.152)	(0.080)	(0.094)
Year Dummies	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1527	1527	1527	1527	1527	1527
Hospitals	324	324	324	324	324	324
Instruments	41	49	65	41	49	65
AR(1) (p-value)	(0.000)	(0.000)	(0.000)	(0.027)	(0.004)	(0.005)
AR(2) (p-value)	(0.039)	(0.094)	(0.205)	(0.453)	(0.459)	(0.461)
Hansen test of overid. (p-value)	(0.443)	(0.083)	(0.238)	(0.090)	(0.060)	(0.053)
Diff.-in-Hansen test of exogeneity (p-value)	(0.524)	(0.125)	(0.286)	(0.640)	(0.304)	(0.229)

Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: The results are based on a system GMM model (Arellano and Bond, 1991; Blundell and Bond, 1998) estimated as in Equation (2.11). Standard errors are corrected for heteroskedasticity. AR(1) and AR(2) are tests for first-order and second-order serial correlation in the first-differenced residuals, under the null of no serial correlation. The Hansen test of overidentification is under the null that all instruments are valid. The difference-in-Hansen test of exogeneity is under the null that instruments used for the equations in levels are exogenous.

the level of practice variation for a focal hospital.

Similar to the mixed (or insignificant) results of previous research on the relationship between hospital quality measures and total cost (e.g., Nair et al., 2013), none of the quality measures are significant as shown in M4, M5, and M6 of Table 2.3. This result is perhaps due to confounding data aggregation factors, inherent in the construction of the explanatory measures and total cost, which possibly cancel out existing effects. We only find a weak association between the level of practice variation and total cost ($\beta = 0.233, p < 0.10$). However, by delving deeper into practice variation at a lower level, particularly for the test-ordering practice, we find an interesting impact on care-delivery cost that we discuss in the next section.

2.4.3 Impact of Variation in Test-Ordering Practice on Care-Delivery Cost

Laboratory/Radiology test-ordering practice is important to assign correct patient diagnoses in a timely manner and to monitor a patient's disease during a care-episode. For each inpatient in our data, we have a detailed list of charges pertaining to a hospital revenue center. A few hundred revenue codes can be categorized as listed in Table A.9 (in Appendix A.9). Almost all patients for the three medical conditions receive laboratory and radiology tests. The amount of charges related to the tests accounts for a considerable portion of a patient's total charges. For example, 98.63% and 97.56% of patients with heart failure received at least one laboratory test and one radiology test, respectively, and on average they account for 27.96% of total charges. Following Equations (2.2) and (2.3), we operationalize measures for underuse- and overuse-variation in test-ordering practice.

To construct dependent variables for examining H2, H3a, and H3b, we obtain *test-ordering charges* by summing charges related to laboratory and radiology tests. Similarly, we sum charges related to medical & surgical supply, cardiology, respiratory, intensive care, coronary care, pharmacy care services, and room charge to obtain *care-delivery charges*. Then, we obtain cost accordingly by multiplying the cost-to-charge ratio of each hospital to the charges.[19]

The results are listed in Table 2.4. M1 to M6 are also included, without a formal hypothesis,

[19]Here we exclude emergency room charge in calculation of either test-ordering cost or care-delivery cost. That is because emergency room charge occurs, in general, simultaneously or prior to (not after) the test-ordering practice.

Table 2.4: Hospital-Level Results of System GMM (WACVGc, Dep: Total Cost, Test Cost, Care Cost)

Dep Var	**Total Cost**			**Cost of Test-Ordering**			**Cost of Care-Delivery**		
	(M1)	(M2)	(M3)	(M4)	(M5)	(M6)	(M7)	(M8)	(M9)
ProcQual (PQ)	0.001	−0.000	−0.004	0.005**	0.004*	−0.029	−0.002	−0.004	−0.035
	(0.004)	(0.005)	(0.017)	(0.002)	(0.002)	(0.023)	(0.003)	(0.004)	(0.056)
ExpQual (EQ)	0.022	0.017	−0.168*	0.042	0.023	−0.221	0.004	0.009	−0.476**
	(0.034)	(0.051)	(0.091)	(0.033)	(0.032)	(0.174)	(0.027)	(0.029)	(0.227)
WACVG_Underuse		0.479**	0.560*		−0.440*	−0.288**		0.406*	0.640***
		(0.208)	(0.290)		(0.234)	(0.144)		(0.240)	(0.212)
WACVG_Overuse		0.109	0.091		0.497**	0.258**		−0.252	−0.203
		(0.068)	(0.077)		(0.197)	(0.126)		(0.177)	(0.206)
PQ*WACVG_Underuse			−0.230			0.111			0.107
			(0.166)			(0.074)			(0.177)
PQ*WACVG_Overuse			0.009			−0.114			0.275
			(0.034)			(0.095)			(0.181)
EQ*WACVG_Underuse			0.299			0.727			1.588**
			(0.671)			(0.527)			(0.729)
EQ*WACVG_Overuse			0.353**			0.165			1.433***
			(0.174)			(0.270)			(0.553)
Teaching Intensity	0.015	−0.002	0.002	−0.014	0.029	0.012	0.037	0.031	0.024
	(0.018)	(0.026)	(0.025)	(0.015)	(0.019)	(0.016)	(0.031)	(0.022)	(0.032)
Bed Size	−0.004	−0.005	−0.008	−0.002	0.003	−0.006	−0.007	−0.006	−0.011
	(0.006)	(0.006)	(0.006)	(0.005)	(0.005)	(0.004)	(0.005)	(0.006)	(0.008)
Case Mix Index	0.000	0.001	−0.001	−0.015	−0.007	−0.008	0.008	0.008	−0.008
	(0.016)	(0.018)	(0.017)	(0.014)	(0.014)	(0.014)	(0.014)	(0.015)	(0.020)
Wage Index	0.146**	0.159***	0.152***	0.015	0.032*	0.027*	0.129**	0.174***	0.231***
	(0.062)	(0.057)	(0.050)	(0.016)	(0.016)	(0.015)	(0.050)	(0.058)	(0.065)
OPDSH Adj Factor	−0.084***	−0.092***	−0.077**	−0.041**	−0.065***	−0.028	−0.081***	−0.101***	−0.061*
	(0.031)	(0.034)	(0.030)	(0.018)	(0.021)	(0.017)	(0.024)	(0.029)	(0.036)
Outlier Adj Factor	0.092	0.134	0.130	0.043	0.049	0.025	0.026	0.053	0.088
	(0.087)	(0.099)	(0.084)	(0.058)	(0.065)	(0.051)	(0.062)	(0.072)	(0.088)
Dep Var$_{(t-1)}$	0.764***	0.740***	0.758***	0.903***	0.890***	0.931***	0.824***	0.774***	0.729***
	(0.087)	(0.080)	(0.068)	(0.029)	(0.032)	(0.032)	(0.058)	(0.066)	(0.071)
Year Dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1519	1519	1519	1519	1519	1519	1519	1519	1519
Hospitals	324	324	324	324	324	324	324	324	324
Instruments	35	51	83	35	51	83	35	51	83
AR(1) (p-value)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
AR(2) (p-value)	(0.264)	(0.254)	(0.240)	(0.588)	(0.547)	(0.571)	(0.425)	(0.481)	(0.977)
Hansen test of overid. (p-value)	(0.404)	(0.321)	(0.201)	(0.101)	(0.077)	(0.167)	(0.278)	(0.412)	(0.083)
Diff.-in-Hansen test of exogeneity (p-value)	(0.127)	(0.265)	(0.376)	(0.148)	(0.281)	(0.387)	(0.389)	(0.307)	(0.340)

Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: The results are based on a system GMM model (Arellano and Bond, 1991; Blundell and Bond, 1998) estimated as in Equation (2.11). Standard errors are corrected for heteroskedasticity. AR(1) and AR(2) are tests for first-order and second-order serial correlation in the first-differenced residuals, under the null of no serial correlation. The Hansen test of overidentification is under the null that all instruments are valid. The difference-in-Hansen test of exogeneity is under the null that instruments used for the equations in levels are exogenous.

to provide insights on how the underuse- and overuse-variations in test-ordering practice relate to total cost (M1 to M3) or test-ordering cost (M4 to M6). For each dependent variable, we first estimate a model with quality measures with controls (as in M1, M4, and M7), add underuse/overuse practice variation measures (as in M2, M5, and M8), and then add interaction terms (as in M3, M6, and M9). Note that, in these models, we include overuse variation in test-ordering practice (i.e., WACVG_Overuse) after controlling for that of underuse practice variation to orthogonalize the two variables, because they are correlated with each other. M3 shows weak evidence that underuse variation in test-ordering practice is positively associated with total cost ($\beta = 0.560, p < 0.10$).

H2 posits that higher underuse variation in test-ordering practice relates to higher subsequent care-delivery cost. Before discussing the results on this hypothesis, we observe from M6 that higher underuse variation in test-ordering practice is associated with lower test-ordering cost ($\beta = -0.288, p < 0.05$). In contrast, overuse variation is positively associated with the test-ordering cost ($\beta = 0.258, p < 0.05$), which make sense from the construction process of the practice variation measures. For the care-delivery cost, which is our main interest, M9 shows underuse variation in test-ordering practice is positively and significantly associated ($\beta = 0.640, p < 0.01$). Thus, H2 is supported. Hospitals that tend to underuse test-ordering practice may spend even more during the care-delivery stages, as we hypothesized.

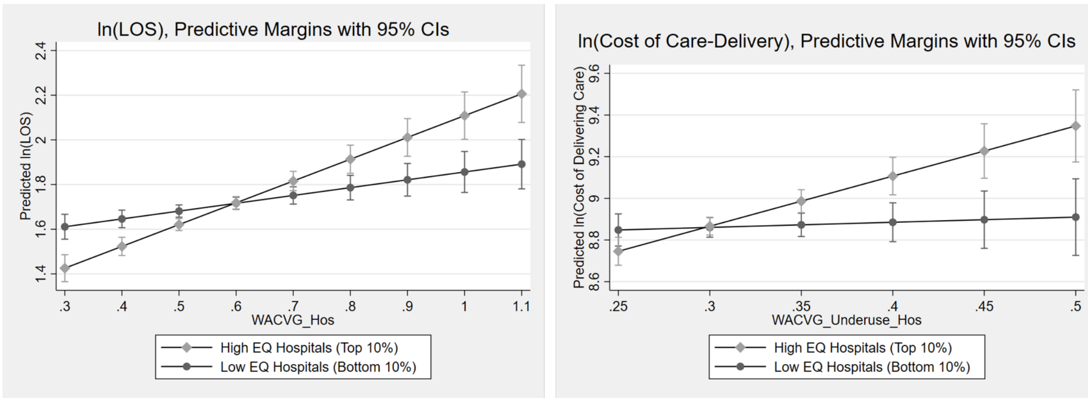
As shown in M9, the interaction between underuse practice variation and process quality is insignificant with care-delivery cost ($\beta = 0.107, p > 0.10$), indicating no support for H3a. However, the interaction between underuse practice variation and experiential quality has a positive association ($\beta = 1.588, p < 0.05$) with care-delivery cost, supporting H3b. Taken together, our findings suggest that high underuse variation in test-ordering practice may lead to higher care-delivery cost, and this relationship is even stronger if a hospital has achieved a high experiential quality measure. As argued earlier, hospitals with both higher underuse variation in test-ordering practice and high experiential quality may have more cases of resource-intensive care reflecting high responsiveness and a tendency to accommodate patient requests, leading to higher subsequent care-delivery costs. We summarize the results of hypotheses tests in Table 2.5.

The left plot in Figure 2.9 represents the interaction between overall practice variation and experiential quality with regard to total LOS. The importance of considering the practice variation in reducing patient LOS is reflected in this plot. Consider the hospitals with high experiential quality (i.e., top 10th percentile). In this case, a 0.1 decrease in overall practice variation would correspond on average to 0.86 days decrease in patient LOS. In contrast, for hospitals with relatively low levels of experiential quality (i.e., 90th percentile), a 0.1 decrease in overall practice variation would result on average in 0.24 days decrease in patient LOS. Similarly, the right plot in Figure 2.9 shows the interaction between underuse variation in test-ordering practice and experiential quality with regard to care-delivery cost. Again, consider the hospitals with high experiential quality. A 0.1 decrease in underuse practice variation in test-ordering would correspond on average to $2150 decrease in care-delivery cost. In contrast, for hospitals with low experiential quality, a 0.1 decrease in underuse practice variation in test-ordering would result on average in only $293 decreases in care-delivery cost. In sum, we highlight the importance of addressing the relationship between practice variation and operational performance, together with hospitals' experiential quality.

Table 2.5: Summary of Hypotheses Testing Results

Hypotheses		Empirical Support
H1:	Higher overall clinical practice variation is associated with higher LOS and total cost.	Supported
H2:	Higher underuse variation in test-ordering practice (e.g., laboratory/radiology tests) relates to higher subsequent care-delivery cost.	Supported
H3a:	For hospitals with higher process quality (PQ), the positive relationships in H1 and H2 are weaker.	Not Supported
H3b:	For hospitals with higher experiential quality (EQ), the positive relationships in H1 and H2 are stronger.	Supported

Figure 2.9: Effect of Practice Variation on Patient Length-of-Stay (Left) and Effect of Lab/Radiology Test Underuse Risk on Subsequent Care-Delivery Cost (Right)



Note: The 10th-90th percentile ranges are displayed for Experiential Quality.

2.4.4 Robustness Checks

Our results remain robust to several checks. First, we construct the measure of practice variation in a different way. In our main analysis, we relied on MS-DRG codes to define groups of patients who require relatively homogeneous care. The MS-DRGs are generally assigned to patients based on multi-dimensional information such as their principal diagnosis and additional diagnoses, the principal procedure and additional procedures, sex, and discharge status. Since the DRG code is assigned around the time of discharge, the MS-DRG based practice variation is an *ex-post* type of practice variation (we named this variable WACVG, where G stands for general). Alternatively, we consider an *ex-ante* type of practice variation. That is, we can rely on the principal diagnosis code, which is usually determined at an early stage of the diagnostic phase, to define a group of patients in calculating practice variation (we named this variable WACVD, where D stands for diagnosis). WACVD, therefore, measures the practice variation for patients who initially got the same main diagnosis but possibly ended up with different clinical pathways. As shown in Appendix A.9, our findings are consistent and robust to alternative practice variation in terms of the point of time for defining of the patient cohort.

Second, in calculating the underuse/overuse variation of test-ordering practice, we use different target standards. Specifically, "County-average" and "State/CBSA[20]-average" could also be worthwhile standards. Indeed, a recent mandatory bundled payment program operated by CMS, namely the Comprehensive Care for Joint Replacement (CJP) Model, adopts a similar approach. In the CJP program, the target price for a care-episode bundle is set prospectively and reflects a blend of hospital-specific and regional data, [21] implying possibilities of using a target standard obtained from a broader regional boundary than a single hospital. Again, as shown in Appendix A.9, our findings are consistent to different target standards.

Third, we test our hypotheses condition-by-condition and consistently find supporting evidence (see Appendix A.10). Fourth, to capture the burden of measuring and reporting process quality that is being updated every year, as listed in Table A.6, we measure the process quality burden (namely, PQB) to address any association between the burden of quality measurement and its impact on operational performance. We find consistent results even after replacing the original PQ measure with the PQB measure, as shown in Appendix A.11.

In sum, our results provide significant evidence to support Hypothesis 1, Hypothesis 2, Hypothesis 3b, but not Hypothesis 3a. Aligned with Salzarulo et al.'s (2011) finding that laboratory/radiology services have the largest impact on the physician's idle time, overall practice variation inside a hospital is associated with longer patient LOS (H1), and this relationship is even stronger when experiential quality is high (H3b). By delving into a more granular level of hospital data, we find that higher underuse variation in test-ordering practice relates to higher care-delivery cost (H2), especially if the experiential quality is high (H3b). Thus, when designing payment reform programs, it is worthwhile to carefully consider the trade-offs between allowing flexible physician practice and providing a standard set of care to improve hospital operational performance.

2.5 Discussion and Implications

Motivated by the move toward payment reform models, in particular bundled payment programs, this study highlights practice variation, an important but understudied metric for research

[20]Core-Based Statistical Area refers collectively to both metropolitan statistical areas and micropolitan areas.
[21]http://www.singletrackanalytics.com/blog/15-11-23/top-ten-things-you-need-know-now-medicare-cjr-program-final

on healthcare operations and healthcare strategic planning. Using a high-frequency inpatient discharge data set from 387 hospitals in NY and FL states with 1,094,111 observations, our findings corroborate the impacts of practice variation on operational performance: practice variation directly affects patients' average LOS during their care episode, and the CMS quality measurement initiative regarding patient experience intensifies this relationship. Also, hospitals with higher underuse variation in test-ordering practice tend to spend significantly greater amounts on subsequent care-delivery cost per capita, especially when the experiential quality is high.

These findings are robust to alternative approaches in measuring practice variation. The strength of our empirical evidence provides assurance to researchers and healthcare managers that the level of practice variation in a hospital should be as salient a concern as the mean level of process and experiential quality measures, which is thus far the major focus of previous literature.

2.5.1 Theoretical Contributions

We suggest several implications for theory. First, our study exposes the role of practice variation in healthcare strategic planning, particularly the explanations of operational performance. Our findings are crucial because previous healthcare operations management research has neglected practice variation, even though it can have a direct bearing on the financial stability of hospitals, payers, and even governments.

Our study of practice variation's role on operational performance advances the healthcare operations literature, because scholars have paid inadequate attention to how practice variation and quality initiatives can have competing direct effects on performance. We find more noticeable effects of quality measures on performance when the level of practice variation is considered together. This result provides a more complex perspective on the translation of quality improvement efforts into better performance and indicates the potential for opposing relationships among the performance drivers.

Furthermore, the interaction effects between practice variation and quality metrics render more nuanced evidence on the role of practice variation in generating better performance. The operational performance metrics (e.g., LOS and care-delivery cost) in our study are more likely to improve for hospitals that reduce practice variation. Hence, there appear to be more dynamic and intricate relationships between quality initiatives (in their mean values) and hospital performance than those identified in previous research. Indeed, for research on the effect of healthcare quality attributes, our findings highlight the importance of including practice variation and its interplay with mandated quality metrics. Taken together, our study supports efforts to move beyond volume or mean and incorporate second-moment information like the variation (we briefly discuss in Appendix A.15 the value of using a dispersion measure in research).

As an interesting aside, CMI, which researchers often use to measure hospital demand uncertainty, ends up insignificant in most of our analyses, unlike in previous studies that do not control for within-hospital practice variation (e.g., Ding, 2014; Senot et al., 2015). This finding possibly indicates that once within-hospital practice variation is taken into account, external demand uncertainty as reflected by CMI may not be such a salient driver of hospital performance.

Practice variation can indicate either a cross-physician inconsistency in preference that produces different practice styles (Wennberg and Gittelsohn, 1973) or differences in characteristics of the social context that provide incentives for certain options (Westert and Groenewegen, 1999). While both perspectives are prominent, our findings on the determinants of practice variation from a post-hoc analysis (see Appendix A.13) find evidence supporting Westert and Groenewegen's (1999) point that emphasizes institution context rather than personal preference.

2.5.2 Managerial Relevance and Implications to Payment Reform Models

We suggest hospital managers and policy-makers should pay attention to practice variation and adopt it as an evaluation metric. Prudent policy-makers may take advantage of practice variation information. For example, they should not only study hospitals with lower practice variation but should also pay attention to hospitals with higher practice variation to understand harmful practice styles and to promote improvement from a wider spectrum of healthcare providers contained within a hospital. In this regard, our findings indicate the relative importance of hospitals with higher practice variation. Thus, policy-makers should pay close attention to hospitals with high practice variation for more productive performance management. Tracking of practice variation, especially high underuse variation in test-ordering practice, may enhance the predictive validity of models linking quality scores to detailed cost components for patient care.

Indeed, a singular empirical research focusing on mean-level quality scores possibly derives inaccurate and biased assessments of hospitals or other types of healthcare providers. With low practice variation, experiential quality improvement leads to shorter LOS and lower care-delivery cost by providing timely and effective services, thus underestimating the power of quality evaluations if practice variation is ignored. Conversely, with high practice variation, experiential quality's impact on the operational performance can be lower, thus overestimating the power of quality evaluations if practice variation is ignored. In this regard, without considering practice variation, hospitals may be over- (under-) rewarded for quality improvements if there is an increase (decrease) in practice variation. Thus, the design of incentives and penalties for better hospital operations should be adjusted to take such effects into account.

For the patient community, we highlight practice variation as a metric that reflects hospital operations. CMS releases several quality measures regularly to help consumers make informed healthcare decisions (CMS, 2014b). Similarly, better-informed patients with earlier or accurate practice variation information might take advantage of WACV-type measures because hospitals with higher practice variation tend to have longer LOS and higher care-delivery cost, which are not preferred by the patient community. Thus, policy-makers should consider practice variation to be a vital healthcare operations metric and add it to the hospital performance dashboard.

From the viewpoint of a policy-maker, our analysis of practice variation provides insights into the operations of bundled payment programs that aim to provide standardized care delivery services for a predefined fixed cost for each specific episode of care. Considering the many unforeseeable situations in healthcare, using a fixed reimbursement approach that imposes less variability, as in the bundled payment scheme, can be a risky strategy for both hospitals and payers. Indeed, managers need a better understanding of the ways to control practice variation and to successfully operate bundled payment programs. Our findings suggest that in a hospital with a lower level of practice variation, patients may stay a shorter period in the hospital during an episode of care. But, the potential savings are not so simple to achieve.

Instead of bundling all services for an episode of care, i.e., providing a standard set of care as much as possible, we suggest BP policy-makers allow some flexible practice or lenient reimbursement standards (perhaps, similarly as in experimental bundled payment schemes, e.g., CMS BPCI initiative, which rely on retrospective fee-for-service based reimbursement). Our findings imply that this approach may be particularly effective for the laboratory/radiology test-ordering practice. By doing so, hospitals operating under a bundled payment system may not simply try providing fewer tests for diagnosis and monitoring of disease (to meet the fixed-reimbursement goal), which can lead to incorrect clinical pathways resulting in even higher spending during later care-delivery stages. Tactically, the suggested approach with lenient reimbursement for test-ordering (possibly also with a reasonable upper limit) may be more effective ultimately to reduce average total cost per patient in hospitals across the nation. Most literature calls for efforts to reduce wasteful overuse of tests, which is clearly important. Meanwhile, our study highlights the risk of underusing tests, which is also pervasive in practice (Zhi et al., 2013). Our practice variation measure also adds value to bundled payment programs since it enables managers to visualize the status of practice variation for each hospital, or even for each medical condition.

One shortcoming of the present uses of quality measures is that CMS provides process quality measures for only a few medical conditions. Reporting quality metrics for a limited range of care may lead to biased decision-making for patients who are not afflicted by one of the conditions on the limited condition list. To balance quality reporting burdens of healthcare providers, CMS annually updates the list of care quality measures that must be reported by healthcare providers. Process quality measures with overall high performance are removed from the list when CMS considers the majority of healthcare providers across the U.S. to have met the quality goal (Mitchell, 2014). CMS then adds new measures that have more opportunity to be improved. However, this topped-out measurement approach seems like a haphazard process of improvement. Also, it risks the existing processes going out of control again. We carefully point to this approach as a cause, among the others, of insignificant estimates of process quality on performance in our analysis.

While our study focuses on medical conditions for which process quality measures are currently available, our metric of practice variation contributes in that it can be applied to any medical condition, enabling future extension of the quality analysis to other conditions as relevant data become available. Compared to CMI, our practice variation metric also enables appropriate analysis of within-hospital variation. Having such a metric is important to accurately evaluate the performance of payment reform models.

In addition, one should not overlook two important points. First, process standards used in performance management should be valid, in that they must either be self-evident measures of quality or be evidence-based (Lilford et al., 2007). Second, in addition to the validity of the measures, the process standards must also be beneficial to healthcare, since the opportunity cost of improving some processes may exceed the contingent gains (Hayward, 2007; Demirezen et al., 2016).

Overall, our study suggests that healthcare policy should be mindful of the potential negative effects of practice variation and introduce provisions that might help to harness such variation. Meanwhile, because such practice variation also might be linked to innovation and continuous improvement in clinical practices, a careful understanding on any adjustments applied to clinical practices should be addressed together.

2.6 Conclusions and Future Research Directions

Motivated by payment reform models and bundled payment policies that aim to reduce practice variation, we contribute to the literature by precisely measuring practice variation within a hospital and by examining the relationship between practice variation and hospital operational performance. From a theoretical lens of statistical process control, we empirically observe a positive association of practice variation on patient LOS and total cost per capita. In addition, we find differential impacts of underuse variation in laboratory/radiology test-ordering practice on the test-ordering cost itself and the care-delivery cost, especially when the process quality and experiential quality measures are taken into account together. Doing so enables managers and policy-makers to understand conditions on which better performance is achieved. We believe that better understanding of such practice variation can lead to successful operations of payment reform models, the target opportunity of which is to reduce waste by decreasing variabilities in care-delivery processes.

Potential limitations of our study motivate several directions for future research. First, a limited number of conditions are examined from a subset of hospitals in New York and Florida states focusing on Medicare populations. These results require confirmation on larger data sets for other regions and for broader spectrums of patient populations. Second, practice variation in a hospital may occur for reasons personal (e.g., physician's preference) or organizational (e.g., hospital characteristics) to the hospital. Behavioral investigation of the causes of such practice variation is warranted.[22] Third, this study finds a relationship between practice variation and hospital operational performance. Meaningful extensions might involve testing on outcome performance to see whether mitigating practice variation can lower readmission and mortality rates (e.g., Senot et al., 2015), as "The Triple Aim: Care, Health, and Cost" is widely pursued and encouraged by the U.S. government,[23] and also on social efficiency to check the impact of unnecessary variation on social values interrelated with other hospitals (e.g., Greenberg and Campion, 2006).

In conclusion, this study documents novel evidence for the role of practice variation in healthcare operations. Practice variation may lead to poor operational performance, and if not managed well, the practice variation can severely diminish the benefits of quality measurement initiatives. We hope this study promotes research to further explicate this important practice variation metric.

[22]In Appendix A.13, we show descriptive evidence of organizational factors as determinants of practice variation.

[23]In Appendix A.12, we test the impact of practice variation on hospital's readmission rate and mortality rate, but more delicate examination on these relationships are warranted.

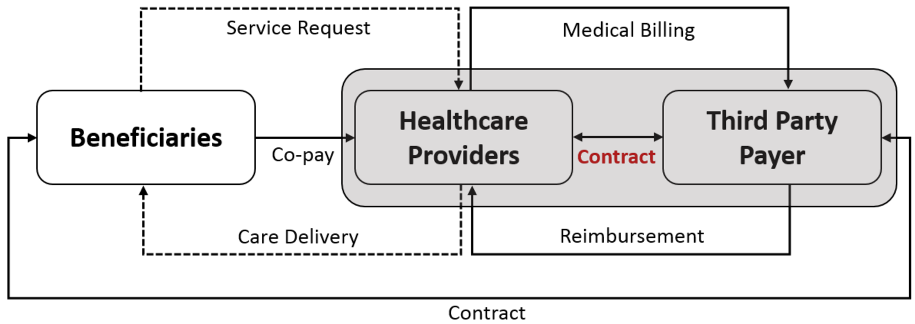
3. SELECTING HEALTHCARE PROVIDERS FOR BUNDLED PAYMENTS IN HEALTHCARE SERVICES

3.1 Introduction

The U.S. healthcare system suffers from high costs and inconsistent quality. Most experts agree that fee-for-service (FFS), which is currently the dominant physician payment system in U.S., is one of the main reasons for increased healthcare cost (Hussey et al., 2011). FFS is a payment model in which each specific service, procedure, or equipment provided is billed and paid for. Under this model, physicians tend to provide more treatments, as payment is directly related to the quantity, not the quality, of treatment. Arrow et al. (2009) also pointed out that the cost and quality problems of the U.S. healthcare system are evident, and recommended to replace the current FFS payment system with a payment system that encourages and rewards innovation in the efficient delivery of quality care. To overcome this problem and to encourage the efficiencies of integrated care, a variety of payment reform efforts such as bundled payments and pay-for-performance have been initiated. A successful changeover requires the development and rigorous evaluation of pilot and demonstration projects that use modified payment mechanisms. The way to select healthcare providers (HPs, e.g., physicians, hospitals, and medical groups) for such experimental projects and even for the nationwide projects in the future is an important issue; however, few studies have explored this issue. This study focuses on how HPs should be selected by policy-makers for a bundled payment program, balancing cost, quality, and efficiency measures. See Figure 3.1 for a schematic of flows of healthcare service delivery.

In what follows, we review two representative payment-reform models: bundled payment model and pay-for-performance payment model, study how they have been applied, and enumerate the challenges associated with these models.

Figure 3.1: Healthcare Service Delivery Flows



3.1.1 Bundled Payment Model (BPM)

Also known as episode-of-care payment or global payment, BPM is the payment of a single price for all of the services needed by a patient for an “episode of care" (Miller, 2009). Such a model is different from FFS where separate payments to HPs are given for each of the individual services they furnish to beneficiaries. Such a system can be used to provide incentives to manage a particular episode of care efficiently with clear accountability. This payment reform model has been tested for several settings. An early evaluation in U.S. was for the Coronary Artery Bypass Draft (CAGB) surgery, which ran from 1991 to 1996 (Cromwell et al., 1997) by the Centers for Medicare and Medicaid Services (CMS). ProvenCare model was developed by Geisinger Health System in Pennsylvania as a bundled payment model for CAGB (Casale et al., 2007). Robert Wood Johnson Foundation experimented with the Prometheus model in which evidence-based case rates were used to decide the total resources to deliver appropriate services for acute and chronic illnesses (Hussey et al., 2011). Currently, CMS has operated Bundled Payments for Care Improvement (BPCI) initiative from 2013 (CMS, 2013a).

3.1.2 Pay-for-Performance (P4P) Model

P4P, also known as value-based purchasing, can be defined as a financial incentive or a payment to the HPs for achieving measurable goals associated with care processes, outcomes, and resources for efficiency and quality (Lindenauer et al., 2007). P4P offers the potential to improve the quality of service delivered, encourages improvement by emphasizing outcomes of care, and also promotes accountability among HPs. In April 2005, CMS implemented its first value-based purchasing pilot named the Medicare Physician Group Practice demonstration (Leavitt, 2006).

3.1.3 Challenges of Payment Reform Models

Payment-reform-models above could effectively motivate HPs to consider the quality of treatment. Gainsharing is one of the financial rewards often discussed, which allows physicians and HPs to receive a share of the savings that result from implementing and coordinating improvements in efficiency and quality. In the case of bundled payments, it increases the transparency and the accountability of treatment cost by grouping related Diagnosis Related Group (DRG) codes for each episode of care. Since the cost is reasonably fixed in advance, bundled payment program is helpful in preventing needless price increases. P4P, on the other hand, rewards physicians, HPs, and medical groups for meeting certain performance measures for quality and efficiency. Therefore, this mechanism prevents physicians from focusing only on the quantity of treatments, in contrast to the FFS system.

The use of bundled payment model for healthcare systems has gained support among healthcare stakeholders as a means to reduce healthcare spending (Hussey et al., 2009) and to encourage coordination across HPs. In addition, quality improvement is expected while discouraging unnecessary care delivery (Miller, 2009).

However, in the early stages of implementing these payment-reform-models, some additional efforts are required. Hussey et al. (2011) discuss conceptual challenges to implement bundled payment models. Participants see value in the bundled payment model, yet the desired benefits take time and considerable effort to materialize. Substantial implementation challenges persist including defining the bundles, defining the payment methods, implementing quality measurement, determining accountability, engaging providers and redesigning episodes of care. Our objective in this paper is to suggest a provider selection model which encourages participation of providers while considering quality measures for the bundled payment model.

3.2 Motivation

This study analyzes opportunities for improvement in the provider selection process used in the previous and ongoing bundled payment programs. Historically, three different selection methods have been used by CMS to select providers: i) Negotiation, ii) Weighted Average Score, and iii) Expert Panel Evaluation based on Relative Weights.

i) Negotiation

CMS used negotiation to select providers in their two bundled payment demonstrations: (a) the Medicare Participating Heart Bypass Center demonstration and (b) the Medicare Cataract Surgery Alternate Payment demonstration. These demonstrations were designed to test the feasibility of a bundled payment for surgical procedures (CMS, 2014c). All the seven participating HPs in the program had significant reductions in total costs (Cromwell et al., 1997, 1998). The method used determines global price for each bundle after several negotiations among all the stakeholders (e.g., physicians, hospital managers, payers, and healthcare consultants).

ii) Weighted Average Score

Evaluation of each HP based on a weighted average score is an alternate selection procedure. Under this system, HPs are scored according to weighted criteria that are used to derive a single metric. For example, in the Medicare ACE demonstration, CMS formed a demonstration review panel and scored applications based on responses to four evaluation criteria, which are: demonstration design (10 %), organizational structure and capabilities (20 %), performance results (35 %), and payment methodology and budget neutrality (35 %). After calculating scores for the criteria, CMS selected higher ranked participants based on the weighted average score (CMS, 2014c). Although this method is easy to understand and utilizes a lot of information in order to make a final decision, it has some limitations. According to McCabe et al. (2005), biases and additive assumptions introduced in the development of the weights and the evaluation can cause problems in the decision-making process. This can be understood as follows. Say, there are two HPs. If one HP suggests an attractive target price of care episode even beyond the mandatory discount rate, then the HP can be a final winner even if its performance is poor compared to the other HP. This decision might not align with the ultimate goal of bundled payment that pursues improved quality with cost reduction. Moreover, it is difficult to conclude that a single weighted average score can consider the flexibility of HPs or their efficiency. Since there are multiple episodes of care included in the demonstration, each provider may prefer to participate in only a subset of the program in isolation from their capability. Also, the HPs may prefer to suggest different target price for a specific episode of care considering the combination of episodes they want to participate in. The overall efficiency of the system can be improved if the provider selection process considers these possibilities.

iii) Expert Panel Evaluation based on Relative Weights

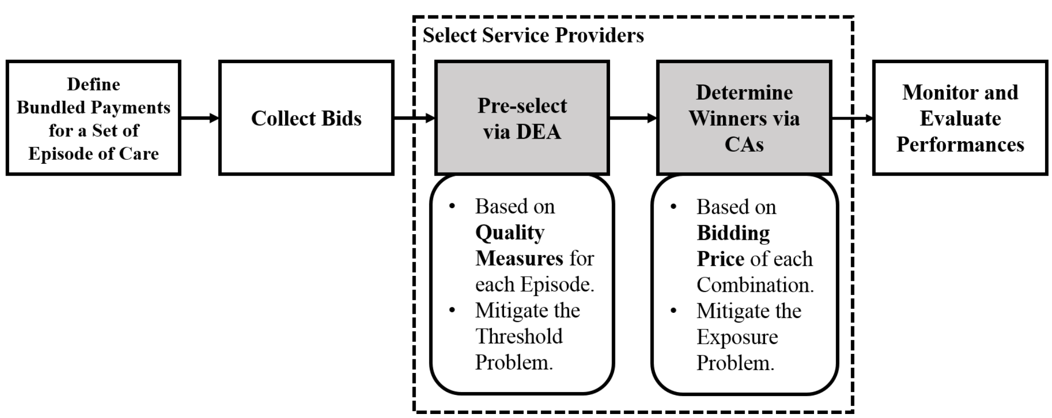
The ongoing BPCI initiative adopts a similar strategy as above and emphasizes evaluation process conducted by an expert panel. Specifically, the CMS screen all applications for eligibility. Each complete and eligible application is reviewed by a panel of experts from the Department of Health and Human Services as well as other governmental and nongovernmental organizations, with expertise in the areas of care improvement, care coordination, and provider payment policy. Reviewed applications will be scored based on the criteria as follows: service model design (20%); financial model (40%); quality of care and patient centeredness (25%); organizational capabilities, prior experience, and readiness (15%). CMS establish guidelines for review panels, and prioritize applications based on both scores and other considerations to select participating HPs (CMS, 2013b). In what follows, we outline the selection method that seeks to overcome potential limitations from the existing selection process.

The focus of this paper is to study the HP selection problem in the context of new and evolving bundled payment proposals for healthcare services. We develop a framework for selecting competitive HPs via a combinatorial auction (CA) with a data envelopment analysis (DEA) as a pre-selection method to evaluate the performance of HPs before the auction process begins.

CA allows bidding of combinations of bundled payment packages. If an auction does not allow a combinatorial bid, then it becomes difficult to exploit synergies of HPs in the various episodes of care. A well designed CA mitigates the *exposure problem* by allowing enough flexility to HPs to express their synergies in the bids. The exposure problem occurs in situations where biddersâĂŹ values are superadditive. In order to win a package that the bidder values more than the sum of the individual items in the package, the bidder might need to bid above her value on the individual items. If the bidder does not end up winning the package, this can expose the bidder to losses. Bidders who are aware of this problem might stop bidding in order to avoid the risk of losses causing low efficiencies and seller revenue (Kwasnica et al., 2005).

However CAs can introduce the *threshold problem* by allowing flexibility in forming the combinations of bundle payment packages. It may favour the bidders who offer attractive discounts even in the absence of synergies within the packages. This in turn affects the efficiency and optimality of the auction. The threshold problem occurs when a number of bidders for small packages must coordinate their efforts to unseat a bidder for a big package. In this case, each bidder has the incentive to allow the other bidders to be the ones who increase their bid in order to displace the big bidder. In principle, all bidders may fail to raise their bids, allowing a particular bid package to win even if it should not have (Kwasnica et al., 2005). To mitigate the threshold problem, we may pre-screen the HPs based on the DEA efficiency score and quality measure before allowing them to participate in the auction process. Figure 3.2 shows the proposed framework of HPs selection procedure for a bundled payment model.

Figure 3.2: The Framework of HPs Selection for a Bundled Payment Model



In sum, our objective is to provide a tool based on DEA and CA to allow a auctioneer (e.g., CMS or private third party payers (TPPs)) to select HPs who will be the participants of bundled payment program. In the pre-selection phase, we evaluate the performance of each HP in terms of efficiency and quality score and narrow down the number of candidates. In the CA phase, we finally determine winners after taking into account various constraints and accommodating the preferences of bidders who have survived from the pre-selection phase. Compared to the existing method that selects high ranked participants based on a weighted average score, the main contribution of our paper is in providing a more structured procedure while considering all the given information, and allowing flexibility for each HP to bid any combination of episodes of care they desire.

3.3 Literature Review

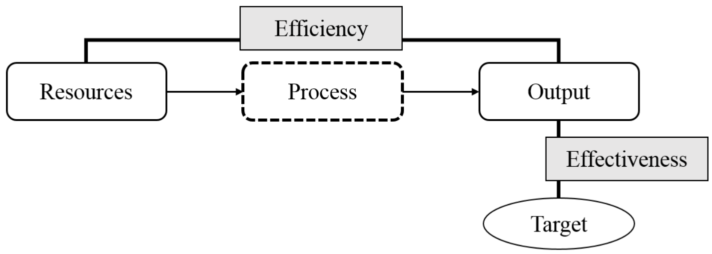
There are three domains of research that are relevant to our work: data envelopment analysis, combinatorial auctions, and evaluation of payment-reform-models. We briefly review them and outline our contributions.

3.3.1 Data Envelopment Analysis (DEA)

DEA is a non-parametric method for the estimation of production frontier and the resulting relative efficiency score. DEA is widely used in healthcare area because it is applicable to the multiple inputs and outputs setting that resembles the nature of the system (Hollingsworth et al., 1999). Compared to the efficiency of manufacturing industries, the measure of service industries such as healthcare or hotel is difficult to define because there is no material output. However, a specification of service efficiency can be realized through the measurement of service outputs.

In addition to the efficiency score, it is critical to adequately deal with effectiveness (i.e., quality) measures. The efficiency and quality measures are the two main components of the performance of HPs (Ozcan, 2008). As shown in Figure 3.3, effectiveness measures how much the provider's targets were reached. They relate to the difference between the actual and the expected values. Typical examples in healthcare sector are 30-day mortality rate, 30-day readmission rate, patient satisfaction score, and so on. On the other hand, efficiency measures how well the resources (such as people, machines, and money) were used to produced output (such as products, services and profit).

Figure 3.3: Performance Indicators: Efficiency and Effectiveness



Recall that the ultimate goal of the bundled payment program is to achieve higher quality and more coordinated care for beneficiaries at a lower cost to the third party payer. Thus, we consider the overall quality level of each HP included in the bundled payment program. Many studies exist that suggest methods to deal with quality measures and DEA models together, and provide insights for the selection framework. Sherman and Zhu (2006) demonstrate that simply adding quality variables as an additional output into the standard DEA model does not help in discriminating the performance and may exhibit a quality and efficiency trade off. In healthcare, managers would not welcome a trade off that sacrifices quality for efficiency. The second approach, which avoids such trade offs, is an evaluation of quality and efficiency independently. Sherman and Zhu (2006) introduce quality-adjusted DEA (QA-DEA) which motivates several extensions (such as in Shimshak et al. (2009), Zervopoulos and Palaskas (2011), Brissimis and Zervopoulos (2012), Choi et al. (2013)).

These studies that provide the way to overcome the inherent problems with using quality measures in DEA, guide the development of our framework, wherein, we adopt the independent evaluation of efficiency and quality. We first evaluate each of the HPs on two dimensions, quality and efficiency, which helps us in setting a minimum threshold level for efficiency and quality, and then we use this an input to the CA.

3.3.2 Combinatorial Auctions (CAs)

CA is an auction in which multiple items are handled simultaneously, and preferences for subsets of the items can be expressed by bidders. Narahari and Dayama (2005) introduce the basic terminologies of CA, the important issues concerning the design of CA, and theoretical issues associated with it.

Epstein et al. (2002) construct a single round sealed-bid CA to assign catering contracts for the school system of Chile. They analyze scenarios based on different food structures, demand levels, performance of firms, and limit on number of firms per region. Their model incorporates the performance of the firm into the objective function by linearly reducing the prices the firms bid by the performance indicator of the firm. In case of our selection problem, however, determining linear weights that balance the levels of price and quality is challenging as it can lead to results depending on the weights. Therefore, we decide to build a model that deals with performance measure and biding price separately. We pre-select HPs based on efficiency and quality scores in order to guarantee minimum performance level before proceeding to the CA stage.

Olivares et al. (2012) conduct an empirical investigation of a large-scale CA for the Chilean school meal auction which is motivated from Epstein et al. (2002). Examining the question of which combinations of bidding should be allowed in the auction process, they find that it is effective to consider the combinations that allow firms' cost synergies originating from economies of scale (i.e., when the volume of service provided increase) and economies of density (i.e., when the bid regions are located nearby one another). Goossens et al. (2014) design a CA to allocate space based on the preferences of many potential users and maximize the total rent while complying with municipal and building regulations.

CA can be powerful in selection process because it can reflect preference and capability of each HP. When we review the previous or ongoing bundled payment programs, they contain multiple number of bundles. However, not all the available HPs are eager to or have capabilities to participate in every episode of care covered under the bundled payment program. Some of the HPs might desire to take part in only a subset of the program considering their preferences or specialty. If all-or-nothing policy exists for the program, it is hard to increase participation rate of HPs. Also, a bundled payment organizer (e.g., CMS) might want to choose competitive HPs in terms of both cost and performance while providing maximum flexibility to the providers. In our framework, we allow bidders to express valuations on bundles of services. Therefore, the resulting bids improve economic efficiency (Cramton et al., 2006) by allocating the bundles to those who value them most at a lower auction cost.

3.3.3 Evaluation of Payment-Reform Models

We briefly review other trials for the evaluation of healthcare payment-reform-models. Even though there are a number of observational or descriptive studies related to the bundled payment programs (Casale et al., 2007; Lindenauer et al., 2007; Paulus et al., 2008; Satin and Miles, 2009; de Brantes et al., 2009, etc.), an operational approach to evaluate or develop payment reform model is still a less researched area.

Denoyel et al. (2014) provide a facility selection model oriented towards HPs under reference pricing. In this work, a payer determines a maximum amount paid for a procedure, and patients who select a provider charging more than the amount fixed by the payer pay the difference. They argue that their model leads to cost reduction for payers, increase in quality of care delivery for patients and enhanced visibility for high-value HPs. Compared to this study, our work focuses more on selection procedures in payer's perspective. Gupta and Mehrotra (2015) examine the proposer selection problem with information sharing under the principal-agent framework for BPCI initiative. They focus on analyzing the ongoing BPCI selection mechanism based on a game theoretical model under various conditions. Our goal, in contrast, is to suggest a new and practically applicable selection framework that considers preference of HPs including other constraints such as geographical constraints (i.e., the minimum number of selected HPs in a single region). Also, extant research does not consider HP selection model equipped with threshold level for both the efficiency and quality measure. We utilize CA in our model to determine the final winners in order to encourage high participation rate of HPs and increase flexibility by allowing them to bid any combination of episodes they prefer.

3.4 A Framework for Healthcare Provider Selection

In this section, we describe the details of the selection framework. First, we use DEA to derive the relative efficiency score for each HP. We then run a CA repetitively while modifying the efficiency and quality cutoff value in order to identify feasible and effective solutions. The two methodologies enable us to determine the winners by considering both bidding price for each episode of care (or combinations of them) and relative performance achieved by HPs. To reduce the risk of project failure, screening process such as pre-qualification is usually applied to contractors before tendering. The first and second steps in our framework align with this practice. In the second step, the efficiency scores are calculated based on representative inputs and outputs of each HP via DEA. In what follows, We discuss the general inputs and outputs that can be used to run DEA model. In the third step, we evaluate the overall quality score for each HP which can be derived based on more lower level quality measures. The HPs with efficiency and/or quality scores below the pre-defined threshold level drop out from the model since they are considered as unqualified compared to others even though they might suggest more attractive price with the intention of joining the bundled payment program. In the last step, a CA determines the winners based on the bidding price suggested by the HPs for each episode of care (or combinations of them).

We represent the overview of our framework in Table 3.1.

3.4.1 First Step: Design of the bundled payment and collection of applications

Design of the bundled payment program precedes selection of participants. It includes all the activities from selecting target episode of cares and defining coverage range of each bundle to establishing operational guidelines. In this section, we provide an activity directly linked to the HP selection process: defining the parameters required for CA. Since a payer might want to focus on some target regions (e.g., statewide, nationwide), preliminary information such as the minimum/maximum number of participants for each region and demand for each episode of care is needed. Also, the payer can define minimum capacity for each episode of care, in order to avoid selecting HPs with limited ability. After these parameters are defined, the payer announces the

Table 3.1: Overview of the Selection Framework

Step 1.		**Design the bundled payment and collect the applications.**
	1.1.	Define the variety and the boundary of the bundles that will be included in the program.
	1.2.	Define the parameters for Step 3 (such as $W^k_{r,min}$, $S^k_{r,min}$, $S^k_{r,max}$, $D^k_{r,min}$. See Table 3.3 for the definitions).
	1.3.	Collect the bids from the healthcare providers.
Step 2.		**Evaluate relative efficiency via DEA.**
	2.1.	Define the scope of efficiency measurement (e.g., facility level, department level, or bundled care procedure level).
	2.2.	Select input and output variables.
	2.3.	Determine the proper DEA model (e.g., orientation (input or output) envelopment surface (CCR or BCC), and other extensions).
	2.4.	Run the DEA model.
	2.5.	Assign percent rank for each relative efficiency score.
Step 3.		**Evaluate quality of care.**
	3.1.	Define the scope of quality (e.g., facility level, department level, or bundled care procedure level).
	3.2.	Select quality measures based on the defined scope.
	3.3.	Derive each value for the quality measures.
	3.4.	Assign percent rank for the quality score.
Step 4.		**Run the combinatorial auction (CA).**
	4.1.	Define the refinement level of cutoff interval for the percent rank of efficiency (σ_g) and quality (ϵ_g) (The smaller intervals, the more detailed results).
	4.2.	Determine the cutoff direction (for example, Figure 3.7).
	4.3.	Analyze the CA results.
	4.4.	Finally, select the preferred cutoff level and the resulting winners.

bundled payment program and collects applications from HPs.

3.4.2 Second Step: Evaluation of Efficiency via Data Envelopment Analysis

Thus, we use a input oriented, constant returns to scale (CRS) DEA model. (We provide more details regarding DEA models with output orientation or variable returns to scale in Appendix B.1). Table 3.2 shows the notations used in the CRS DEA model. The objective of the model is to maximize the ratio of weighted multiple outputs to weighted multiple inputs. Any HP compared to others should have an efficiency score of 1 or less, with either 0 or positive weights assigned to the inputs and outputs (Ozcan, 2008).

Table 3.2: Notations used in CRS DEA model (Hospital Level)

Sets, Indices, and Parameters:	
N	Set of healthcare providers.
$i \in N$	Index for a HP.
P	Set of inputs.
$p \in P$	Index for inputs.
Q	Set of outputs.
$q \in Q$	Index for outputs.
I_{pi}	Value of input p parameter for HP i.
O_{qi}	Value of output q parameter for HP i.
Decision Variables:	
v_p	Weights for the inputs.
u_q	Weights for the outputs.

$$\underset{v,u}{\text{Maximize}} \qquad \frac{\sum_{q\in Q} u_q O_{qo}}{\sum_{p\in P} v_p I_{po}}$$

Subject to

$$\frac{\sum_{q\in Q} u_q O_{qi}}{\sum_{p\in P} v_p I_{pi}} \leq 1, \forall i \in N, \tag{3.1}$$

$$\frac{u_q}{\sum_{p\in P} v_p I_{pi}} \geq \epsilon, \forall q \in Q, \tag{3.2}$$

$$\frac{v_p}{\sum_{p\in P} v_p I_{pi}} \geq \epsilon, \forall p \in p \tag{3.3}$$

$$v_p, u_q \geq 0$$

The subscript $o \in N$ in the objective function denotes a focal decision making unit (DMU). Each HP, in turn, becomes a focal HP when its efficiency score is being computed. Constraint (3.1) ensure that no HP is more than 100 percent efficient. Constraints (3.2) and (3.3) ensure that each weight for the inputs and outputs is strictly positive. After we derive the efficiency scores, we assign the percent rank for each HP. For example, if the efficiency score of HP A is the 10^{th} highest value among 200 HPs, then the percent rank of HP A becomes 0.05. Based on these percent rank values, we remove HPs with efficiency below the pre-defined threshold percent rank level. (We explain why we use percent rank instead of raw score in Section 3.7.2).

3.4.3 Third Step: Evaluation of Effectiveness (Quality)

In the next step, we evaluate the quality score of each HP. An example for the HP level quality measure is readily available from CMS database as *Total Performance Score (TPS)*. TPS is a composite measure capturing hospital quality performance related to clinical process performance (45%), patient experience (30%), and outcome performance (25%) based on the Hospital Value-Based Purchasing (HVBP) Program, which is a part of CMS's to link quality into Medicare's payment system (CMS, 2014a). The measure incorporates both quality attainment and quality improvement. Higher scores indicate higher quality performance.

Alternatively, instead of hospital-level overall quality score such as TPS, quality measures for each episode of care can be considered. The quality measures might vary across the episodes of care. For example, in case of surgical type Medicare Severity (MS)-DRG, readmission rate and mortality rate is frequently considered, while readmission rate and complication score is considered for medical type of MS-DRG. Thus, in this approach, quality scores (Q_i) should be measured separately for each episode of care.

After we derive the quality score, similar to the efficiency score, we assign the percent rank for each HP. Based on these percent rank values, we remove HPs with low quality below the pre-defined threshold percent rank level.

Table 3.3: Notations used in Combinatorial Auction

Sets, Indices, and Parameters:	
K	Set of episodes of care to be provided.
$k \in K$	Index for an episode of care.
R	Set of Regions (for example, Public Health Region (PHR)).
$r \in R$	Index for a Region.
l_{ir}	Binary parameter which takes value 1 iff HP i is located in region r.
c_i^k	Volume capacity for episode k of HP i (per year).
M	Set of bids allowed for a HP.
$j \in M$	Index for a bid.
B_{ij}	Zero-one vector composed of $a_{ij}^k, k = 1, ..., K$.
a_{ij}^k	Binary parameter which takes value 1 iff bid j of HP i will provide the entire service corresponding to episode k.
p_{ij}	Bundled price associated with the bid B_{ij}.
$\bar{p}_{ij}$	Historical average total payment associated with the bid B_{ij}.
$S_{r,min}^k$	Minimum number of winners required for episode k in region r.
$S_{r,max}^k$	Maximum number of winners required for episode k in region r.
$W_{r,min}^k$	Required minimum volume threshold of episode k in region r for each HP to be selected.
$D_{r,min}^k$	Required minimum volume episode k in region r (i.e., demand).
E_i	Efficiency score percent rank for each HP i.
Q_i	Quality score percent rank for each HP i.
G	Set of subset ID.
$g \in G$	Index for an subset ID.
σ_g	Threshold value for efficiency score percent rank of subset ID g.
ϵ_g	Threshold value for quality score percent rank of subset ID g.
Decision Variables:	
x_{ij}	If $x_{ij} = 1$, then the bid B_{ij} is selected, Otherwise, $x_{ij} = 0$.
y_i	If $y_i = 1$, then the HP i is selected, Otherwise, $y_i = 0$.

3.4.4 Fourth Step: Determination of Winners via Combinatorial Auction

Finally, we run a CA to determine winner HPs based on discounted price for each bundle suggested by the HPs. Additional sets, indices, parameters, and decision variables used in the CA model are listed in Table 3.3.

$$\text{Maximize} \quad \sum_{i \in N} \sum_{j \in M} (\bar{p}_{ij} - p_{ij}) x_{ij}$$

Subject to

$$\sum_{i \in N} \sum_{j \in M} a_{ij}^k x_{ij} \geq 1, \quad \forall\, k \in K, \tag{3.4}$$

$$\sum_{j \in M} x_{ij} \leq y_i, \quad \forall\, i \in N, \tag{3.5}$$

$$W_{r,min}^k y_i l_{ir} \leq c_i^k, \quad \forall r \in R, \forall k \in K,\ \forall i \in N, \tag{3.6}$$

$$S_{r,min}^k \leq \sum_{i \in N} \sum_{j \in M} l_{ir} a_{ij}^k x_{ij}, \quad \forall r \in R,\ \forall k \in K, \tag{3.7}$$

$$S_{r,max}^k \geq \sum_{i \in N} \sum_{j \in M} l_{ir} a_{ij}^k x_{ij}, \quad \forall r \in R,\ \forall k \in K, \tag{3.8}$$

$$D_{r,min}^k \leq \sum_{i \in N} \sum_{j \in M} l_{ir} a_{ij}^k c_i^k x_{ij}, \quad \forall r \in R,\ \forall k \in K, \tag{3.9}$$

$$E_i y_i \leq \sigma_g, \quad \forall i \in N,\ \forall g \in G \tag{3.10}$$

$$Q_i y_i \leq \epsilon_g, \quad \forall i \in N,\ \forall g \in G \tag{3.11}$$

$$x_{ij} \in \{0, 1\}, \quad \forall i \in N,\ \forall j \in M, \tag{3.12}$$

$$y_i \in \{0, 1\}, \quad \forall i \in N \tag{3.13}$$

The above formulation denotes a mixed-integer-programming for the CA. This step maximizes the discounted amount between the past average total cost under FFS and the bidding price while satisfying requirements such as the minimum volume threshold for each bundle, the minimum and maximum number of HPs for each region, and the demand for each region. Constraint (3.4) states that each episode k should be covered by at least one bid combination. Constraint (3.5) ensures that if no bid from HP i is chosen, then $y_i = 0$, where y_i is the indicator showing whether the

HP i is selected or not. y_i is 1 only if one of x_{ij} becomes 1 among possible bid combinations. Constraint (3.6) states that the capacity of episode k by HP i should be beyond the minimum volume threshold of it. Constraints (3.7) and (3.8) ensure the required minimum and maximum number of HPs should be selected for each episode of care k and for each region r. Constraint (3.9) means at least minimum volume of episode k in region r should be covered by selected HPs. Decision variables should only have binary value by constraints (3.12) and (3.13).

Constraints (3.10) and (3.11) ensure that HPs who have the percent rank value below the threshold values are eliminated. Even though this process is included to guarantee a minimal level of efficiency and quality level and to motivate HPs to manage them well, it is not easy to determine what fraction of HPs can drop out based on efficiency and quality percent rank criteria. This is because the given information (specifications, claim data of HPs, and their bidding price) and other constraints (e.g., minimum and maximum number of HPs for each region, demand) make the problem complex and it is non-trivial to assure the feasibility of the problem. If the cutoff levels are high, the problem becomes infeasible, whereas if the levels are too low, the opportunity to select more competitive HPs may be lost.

We iteratively modify the threshold percent rank values σ_g and ϵ_g where $g \in G$ step-by-step. The intervals of threshold value can be flexibly determined based on the payer's decision. Note that as intervals are more precisely defined, more detailed results for each case can be obtained, but the computation time is longer. The iteration runs as follows. If the payer desires to check the selection tendency with threshold percent rank interval with 0.1, then we first run the CA model with $\sigma = 1$ and $\epsilon = 1$. The results of this first trial becomes first best solution because the model selects the winners while never considering the efficiency or quality factors. Subsequently, we run the model again with $\sigma = 0.9$ and $\epsilon = 0.9$ which means we don't allow the HP whose efficiency percent rank or quality percent rank is included in lower 10%. The steps are repeated until we reach an infeasible status. In Section 3.7, we formulate and provide an simple example of direction for shrinking feasible region. However, the direction and preferentially considered threshold values can be defined based on the payer's decision.

3.5 Problem Settings

The hypothetical bundled payment program assumed in this study is based on practical settings from previous (e.g., ACE Demonstration) or ongoing (e.g., BPCI Initiatives) programs operated by CMS. Payers want to increase the opportunity for the largest number of qualified applicants by concentrating on target regions and to ease the administrative burden of implementation and evaluation (CMS, 2014c). While imitating this principle, we slightly redefine the boundary of region. We assume that the payer aims to cover each region by allowing at least a predefined minimum number of HPs in order to provide easy access to healthcare services for patients. States are usually too large to be considered as a single region when we look at the distance between the two farthest points. If the selected providers are densely located in a small spot of each state, some patients will inevitably spend significant amount of time on the road: this is undesirable. On the other hand, cities or towns are usually too small to be considered as a single region because some smaller cities may not have enough number of providers to compete with each other. Thus, we decided to define region as a sub-area of each state. If we select providers based on this concept of regions, it is possible for almost all patients to reach their nearest provider within a reasonable time. In our selection process performed based on real data, we assume that applicants exist in each of the public health regions (PHR) defined by the Departement of Health and Human Services.

Hospitals with higher volumes of certain surgical procedures are known to have better results, and surgeons who perform more of certain operations have fewer patient deaths (CMS, 2014c). Therefore, applicants for bundled payments are often required to show that they meet specific volume thresholds (as in ACE demonstrations). Table 3.4 shows an example. In addition, every participant in the selction practice can bid on any combination of bundles considering their capability, but not for any portion of a single bundle. We consider these constraints in our CA model.

In the previous demonstration of bundled payment, applications were scored, in part, on the percentage discount across all selected bundles. Discount on current Medicare rates was given a significant weight as part of the evaluation of the overall application. The applicant's overall global bid relative to other proposals received was also considered. Applicants need to provide

Table 3.4: Required Minimum Volume Thresholds for each Episode of Care in ACE Demonstration

Procedures	Medicare	Total
Coronary Artery Bypass Graft (CABG)	100	200
Percutaneous Coronary Intervention (PCI)	200	400
Hip or Knee Replacement (HKR)	90	125

sufficiently competitive discounts to Medicare to yield meaningful savings to both the beneficiary and the Medicare program (CMS, 2014c). We utilize the concept of discounted payment amount suggested by each HP in our selection process. Specifically, this information becomes the bidding price (p_{ij}).

3.6 Input and Output Measures of DEA

Healthcare industry usually provides services rather than physical products. Thus, researchers in healthcare sector have used such measures to derive efficiency scores that reflect their topics and data availability. The measures also align with whether the target scope is a hospital, a physician, or each episode of care. Efficiency can also be measured and assessed for different aspects or segments of care (i.e., episodes of care ranging from management of a condition over time to specific procedures) and across different levels of organizational accountability (e.g. individual physicians, physician organizations, hospitals, insurance plans, or accountable care organizations). Decisions on the appropriate level of measurement and accountability will depend on the purpose of measures. Since bundled payment program comprises various episode of cares to be provided by HPs, either facility-level efficiency or medical condition-level efficiency can be considered. If the scope of efficiency is HP, each provider becomes a decision making unit (DMU) and if the scope of efficiency is medical condition, then each claim data can be considered as DMU. It is notable that the discriminatory power of DEA can be reduced if the number of inputs and outputs are relatively highter than the number of available DMUs. In this study, we focus on the hospital-level[1] DEA model. Banker et al. (1989) suggest a rough rule of thumb: If p is the number of inputs

[1]For the medical condition-level DEA model, refer Appendix B.1. The idea is to calculate relative efficency score for each medical condition. In such a scenario, measures that properly represent inputs and outputs used for the medical condition can be utilized.

and q is the number of outputs used in the analysis, then the sample size n should be greater than equal to $max\{pq, 3(p+q)\}$.

Selecting different input and output variables could influence the results of the DEA model. Indeed, DEA estimates relative efficiencies (i.e., relative to the best practice frontier) and allows for specialization in one or another input or output variable. In what follows, we provide examples of measures in Table 3.5 for the hospital-level DEA model provided by Ozcan (2008), Cooper et al. (2011), and Ozcan and Lynch (1991). The purpose here is to capture the managerial performance that can be attributed to hospital management.

Table 3.5: Hospital-level DEA Input and Output Measures

Input Measures:	
Number of Operational Beds	This is a proxy variable for capital investments.
Service Complexity	The number of diagnostic and special services provided exclusively by the hospital can be used as another capital proxy variable.
Full Time Equivalents (FTEs)	Labor is the second major category for hospital inputs. In evaluating the performance, Ozcan (2008) insists that it is prudent to attribute the labor as non-MD labor or FTEs. The number of non-MD FTEs employed by a hospital would cover all nursing, diagnostic, therapy, clerks and technical personnel.
Other Operational Expenses	This variable provides the account for medical supplies, utilities, etc. to provide the services to patients except capital investments and labor expenses.
Output Measures:	
Case-mix Adjusted Discharges	The number of discharges accounts for inpatient services. Since not all patients arriving at the hospital require same level of attention and service, we account for this diversity in health service demand or its provision by CMS case-mix index[2] for hospitals. The case-mix index is calculated based on DRG codes providing relative weight for acuity of the services provided by a hospital.
Outpatient Visits	One can differentiate the visits based on whether these are day surgery, emergency or routine visits if available.

3.7 Numerical Analysis

In this section, we perform numerical analysis based on real data for HPs located in Texas, USA to illustrate the selection process.

3.7.1 Simulation Settings

To develop mechanism for CA and examine the effects of its parameter values, we pre-define four bundles, as per MS-DRG code 194, 280, 291, and 470: Pneumonia, Acute Myocardial Infarction, Heart Failure, and Major Joint Replacement. The discharged volumes for these MS-DRG codes are large and there are previous and ongoing bundled payment programs that consider these as bundles. We assume that if the HPs have capabilities of care for at least one of these four bundles, then they apply for the bundled payment program and reflect their preference and capabilities in their bidding price.

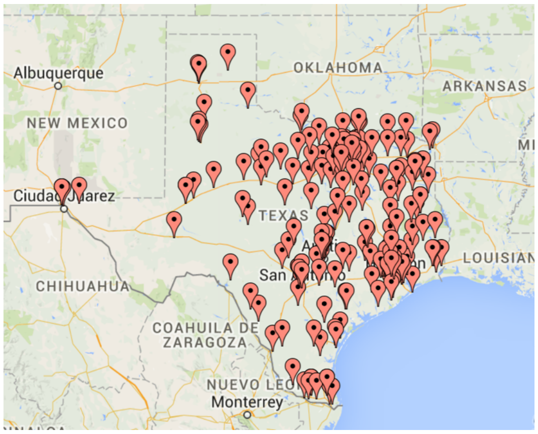
As mentioned earlier, PHR is used as the size and boundary of the region denoted by r. Texas, the target state in this experiment, has 305 HPs and is divided into 11 PHRs (Figure 3.4). Among the 305 HPs, 285 have the capability of treating at least one of the four target bundles, and are categorized into each PHR as in the Table 3.6. The numbers of HPs for each PHR and each bundle are listed in Table 3.7. For each r, the values for CA parameters such as $W^k_{r,min}$, $S^k_{r,min}$, $S^k_{r,max}$, and $D^k_{r,min}$ in Table 3.3 can be determined.

The payer may desire to restrict the number of selected HPs for each region in CA step, in either (or both) minimum or (and) maximum way. We deploy this restriction by multiplying parameters called ***MinWin*** and $MinMax$ to the number of available HPs. For example, if ***MinWin*** $= 0.1$ and $MinMax = 0.3$, then the number of available HPs multiplied by these two parameters provide the minimum and maximum required numbers (i.e., $S^k_{r,min}$, $S^k_{r,max}$) of HPs for each region and each bundle. In a similar way, we set up the other parameter values since we can check the volume for each HP and for each bundle based on the historical FFS records. By multiplying predefined ratio to the historical volume data, we derive $W^k_{r,min}$ and $D^k_{r,min}$.

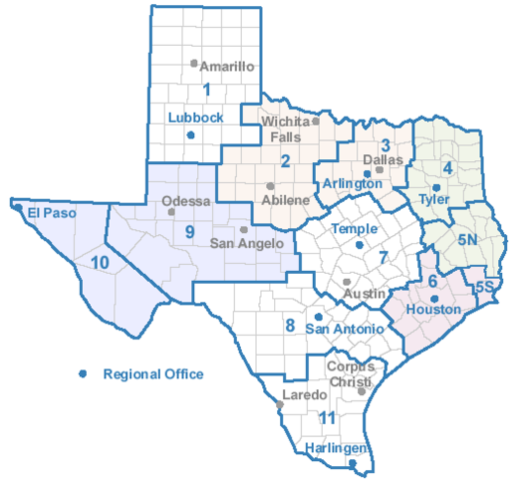
Table 3.6: The Number of HPs for each Public Health Region

PHR	1	2	3	4	5	6	7	8	9	10	11	Total
Num of ProviderID	16	15	77	21	14	53	31	22	9	7	20	285

Figure 3.4: Healthcare Providers in Texas (a), Public Health Regions in Texas (b)



(a)



(b)

Table 3.7: Available Healthcare Providers in each Public Health Region for each Episode of Care Bundle

	PHR1	PHR2	PHR3	PHR4	PHR5	PHR6	PHR7	PHR8	PHR9	PHR10	PHR11
DRG194	13	14	59	20	14	43	29	19	9	6	18
DRG280	5	2	34	10	4	30	15	8	5	3	13
DRG291	6	6	55	16	10	41	25	15	7	4	18
DRG470	8	6	70	14	8	43	24	17	6	6	15

3.7.2 Results and Interpretations

Following the framework, we first derive the efficiency score for each HP by using the measures in Section 3.6 for the HP level efficiency. Among the measures, Case-mix Adjusted Discharges and Outpatients Visits are obtained from *CMS Hospital Compare*[3] database. The data was collated from *Healthcare Information and Management Systems Society (HIMSS) Analytics Database*[4] and/or *American Hospital Association (AHA) Database*. The correlation between the inputs and outputs for DEA are listed in Table 3.8. Histogram of DEA efficiency scores is shown in Figure 3.5. For

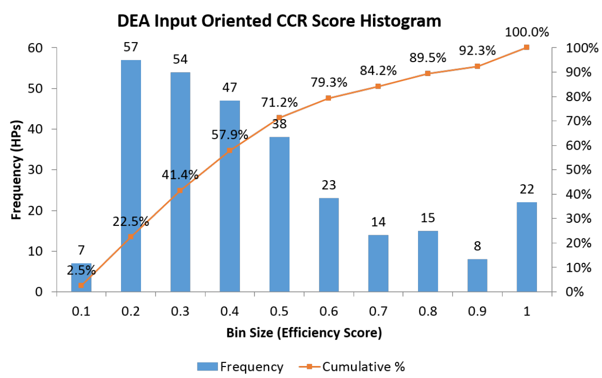
[3]https://data.medicare.gov/
[4]http://apps.himss.org/foundation/histdata.asp

example, there are 8 HPs whose efficiency score is greater than equal to 0.8 and less than 0.9.

Table 3.8: Correlation Table of Variables used in DEA

	(1)	(2)	(3)	(4)	(5)	(6)
(1) Number of Beds	1					
(2) ServiceMix	0.0222	1				
(3) FTEs	0.0664	-0.0192	1			
(4) OtherOpExpense	0.6204	-0.0539	0.1819	1		
(5) CaseMixAdjDischarge	0.4523	0.0011	0.3403	0.4021	1	
(6) OutpatientVisits	0.5604	-0.0468	0.1145	0.7873	0.3854	1

Figure 3.5: Histogram of DEA Score



In case of quality measures, we utilize the TPS score that is available from CMS Hospital Compare database. Figure 3.6(a) shows the plot of the raw efficiency score and the raw quality score in two-dimensional space. We observe that the HPs are distributed densely in some spots (e.g., DEA $\in [0.25, 0.5]$ and TPS $\in [35, 53]$). Since we will repetitively run the CA model while updating cutoff value of DEA score and TPS with fixed interval, it is more reasonable to start from evenly distributed space. This is because, the number of HPs dropped for each iteration highly depends on their distribution on the two-dimensional space. Hence, from a payer's viewpoint, it becomes difficult to define the cutoff intervals to drop a “reasonable" number of HPs in each iteration. To mitigate this problem, we re-plot the data in terms of percent rank rather than raw score as in Figure 3.6(b). In this plot, the HPs in the bottom-left region are high performers for both efficiency and quality, whereas the HPs in the top-right are the possible candidates to be dropped first. The concept of percent rank is also aligned with the overall selection process. In order to cover regional demands, it is customary to consider the required number of HPs based on their rank instead of their absolute DEA or quality score.

After obtaining the two-dimensional percent rank plot, we define the direction and interval of the cutoff values (Note that both direction and cutoff interval can be flexibly defined reflecting the payer’s preference.). One possible direction is shown in Figure 3.7(a). Figure 3.7(b) shows an alternative way that we focus on in this numerical analysis. We take 25% as the cutoff interval of percent ranks; thus there exist four intervals for both efficiency and quality percent rank resulting in 16 subsets of HPs in Figure 3.6(b), namely subset ID k; $k = 1, 2, ..., 16$. In each iteration of CA, the setup in Figure 3.7(b) keeps on rejecting a larger balanced number of HPs corresponding to each subset ID compared to Figure 3.7(a). Also, the direction in Figure 3.7(b) is quality-prioritized in the sense that it tries to drop HPs with poor quality first. For example, after the 4$^{\text{th}}$ iteration of CA (i.e., when we drop subset ID 1, 2, 3, and 4), the lower 25% of HPs with bad quality are completely dropped. Thus, we can execute finite rounds of CA for winner determination with the remaining upper 75% HPs in terms of quality.

Now, we discuss the settings of CA and how it works. Since we assume four different bundles, there exists a total of fifteen ($= 2^4 - 1$) possible combinations of bundles (i.e., *packages*) (Table B.2 in Appendix B.2). Each HP can bid on a single or multiple numbers of packages considering their preferences. However, each HP can be selected as a winner with at most a single package among the packages they bid.

We determine the bidding prices of each package as follows. Average total cost under FFS for each MS-DRG (i.e., bundle in this setting) and for each HP are obtained from CMS database. Descriptive statistics are listed in Table 3.9. Based on the existing data, we calculate the bidding

Figure 3.6: Efficiency vs. Quality Plot for Raw Score (a) and Percent Rank (b)

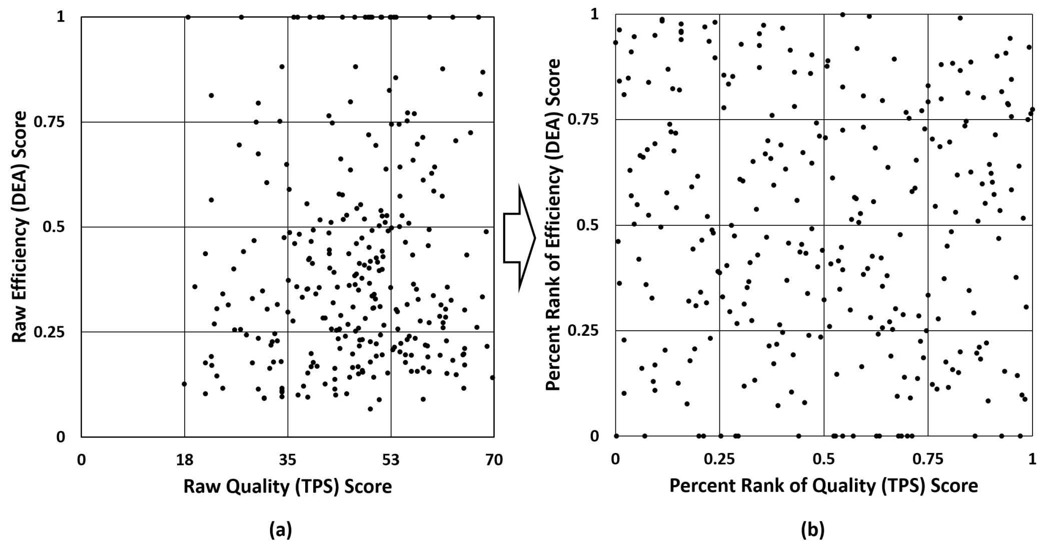
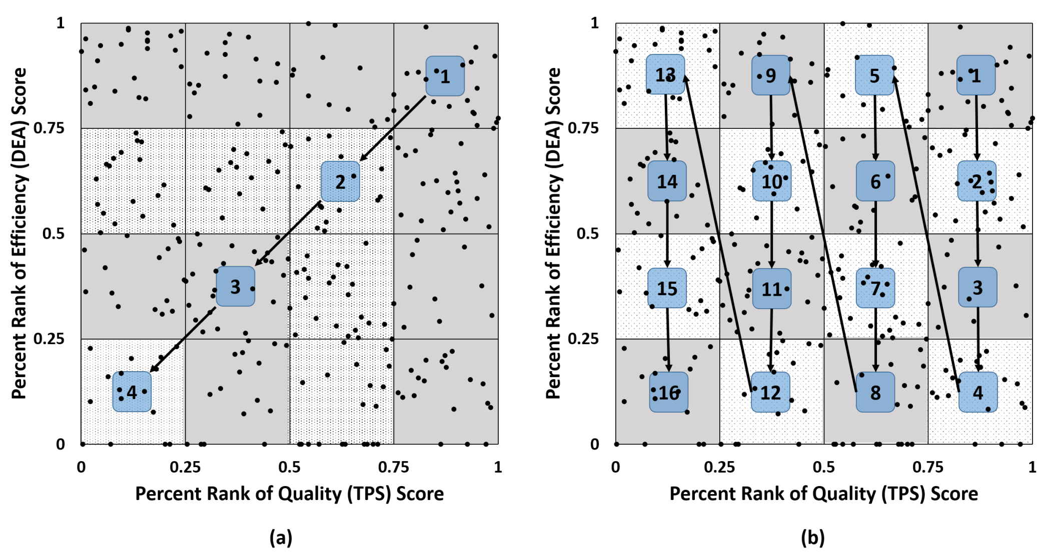


Figure 3.7: Shrinking Feasible Region on a Percent Rank Plot - (a) Basic, (b) Alternative



price as the sum of average total cost of the bundles included in the package multiplied by a random number generated from interval $[0.90, 1.05]$. As, bundled payments pursue cost reduction, the lower limit of 0.9 is used for the interval. However, it is also possible that some HPs might strategically act by bidding some bundles even higher than their historical average value. This is possible if either they really do not want to take care of that package, or rely on the small possibility of getting more reimbursement for the bundles under the decision that they are still competitive compared to others. Table 3.10 provides descriptive statistics regarding care volume data for the four bundles under FFS.

Table 3.9: Descriptive Statistics of Average Total Cost for each Bundle in 2013

Average Total Payment	DRG194	DRG280	DRG291	DRG470
Mean	6,712.42	11,940.20	9,973.86	14,129.19
Standard Deviation	1,352.66	2,448.35	2,081.38	2,360.31
Standard Error	86.6	215.57	146.08	160.23
Median	6,301.34	11,429.72	9,377.21	13,590.55
Minimum	5,125.84	8,071.13	7,530.65	10,355.88
Maximum	16,481.00	21,682.22	21,249.79	25,201.04
# of HP	244	129	203	217

Table 3.10: Descriptive Statistics of Volume for each Bundle in 2013

Volume	DRG194	DRG280	DRG291	DRG470
Mean	53.79	32.20	63.84	126.26
Standard Deviation	40.50	26.52	58.57	136.84
Standard Error	2.59	2.33	4.11	9.29
Median	42	23	44	85
Minimum	11	11	11	11
Maximum	258	152	401	823
Sum	13124	4154	12960	27399
# of HP	244	129	203	217

We run the CA with a wide range of parameter values. For brevity, we include the results with respect to the scenario with $MinWin = 0.1$ and $MaxWin = 0.3$. Figure 3.8 shows the results of objective value, total number of selected HPs, and the selected HPs ratios. The top-right dotted bar in Figure 3.8 shows the first best result, that is, when no HP is dropped before the CA step. In other words, quality and efficiency level of HPs are not considered. When HPs in subset ID 1 (i.e., percent rank of DEA in $[0.75, 1]$ and percent rank of TPS in $[0.75, 1]$) are dropped, the objective value decreases ($\$127,336.67 \rightarrow \$126,397.10$). Note that subset ID is linked to that in Figure 3.7(b). Other results can be understood in similar way. Also, when the HPs in subset 12 are dropped, the CA cannot derive any feasible solution, thus the process terminates. We provide another set of results with $\mathit{MinWin} = 0.4$ and $\mathit{MaxWin} = 0.7$ in Figure 3.9. Obviously, this scenario requires more number of HPs to be selected in each region compared to the previous scenario, and therefore terminates quickly compared to the case in Figure 3.8.

Let us explore the scenario with $\mathit{MinWin} = 0.1$ and $\mathit{MaxWin} = 0.3$ a little further. As shown in Figure 3.10, the optimal objective value of CA model is non-increasing as HPs are dropped. Meanwhile, the ratio of selected HPs to available HPs increases as more HPs are dropped. Note from Figure 3.11 that the number of available HPs decreases linearly as we reduce the feasible region, whereas the number of selected HPs fluctuates around a mean.

Finally, we examine the relationship between maximum discounted amount (i.e., maximum cost reduction) and quality/efficiency level. Since it is possible that HPs with poor quality level may suggest better bid prices, we need to control for the same. A clear trade-off between quality and cost reduction is visible in Figure 3.12. Specifically, the average TPS score and the minimim TPS score among selected HPs are increasing while the maximum cost reduction is decreasing as more HPs are dropped. In the case of cost reduction versus efficiency level, we see a weak trade-off as shown in Figure 3.13. The reason why the minimum DEA score remains constant is due to the direction of cutoff values we adopted (as shown in Figure 3.7). Since the HPs with lower DEA score survive in the dataset even until the selection process terminates, if some of HPs with lower DEA score are selected, then there is little room for improvement of minimum DEA score. This is due to the constraint (i.e., Equation (3.10)) in CA.

This selection process lets the CMS (who will operate the bundled payments) have flexible options to choose from. For example, if they emphasize the importance of quality level, then they may choose the selected HPs from the iteration of subset ID 8, 9, 10, or 11 in Figure 3.12 while having some cost reductions. On the other hand, if budget constraints matter, then they may consider the options of subset ID 4, 5, 6, or 7 as alternatives. In sum, the payer can selectively decide winner groups while balancing bid prices of HPs and their performance. Lastly, we visualize a possible winner determination result as shown in Figure 3.14.

Figure 3.8: Selection Results with $MinWin = 0.1$ and $MaxWin = 0.3$

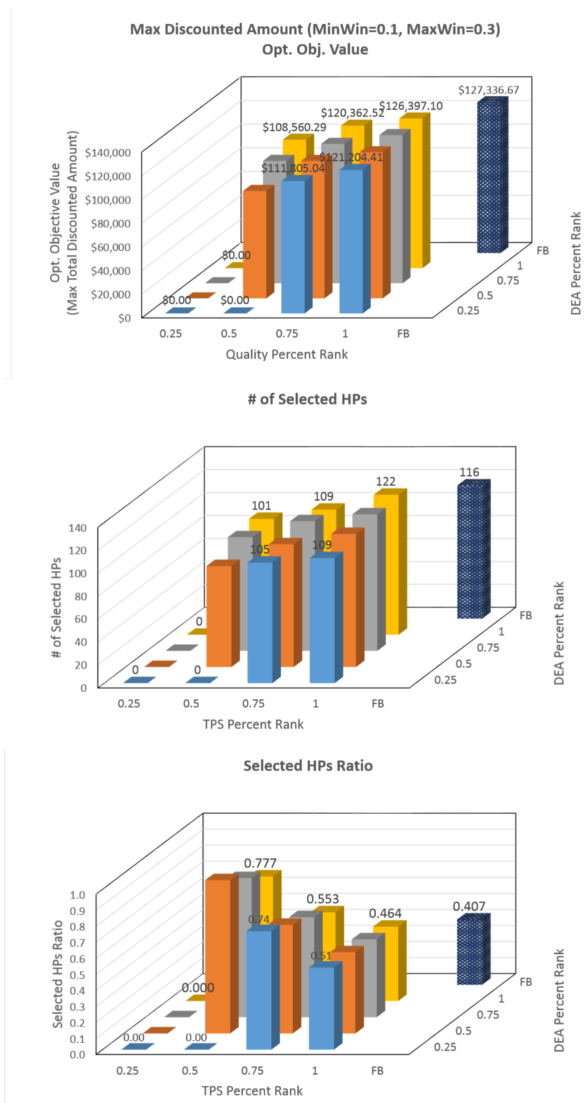


Figure 3.9: Selection Results with $MinWin = 0.4$ and $MaxWin = 0.7$

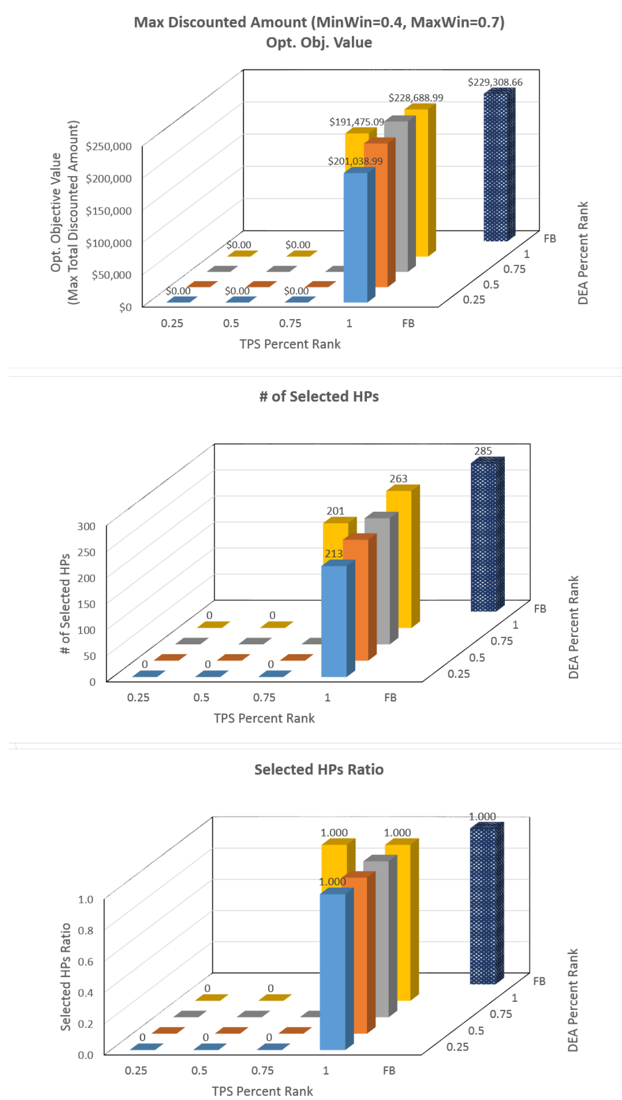


Figure 3.10: Objective Function Value vs. Selected HP Ratio

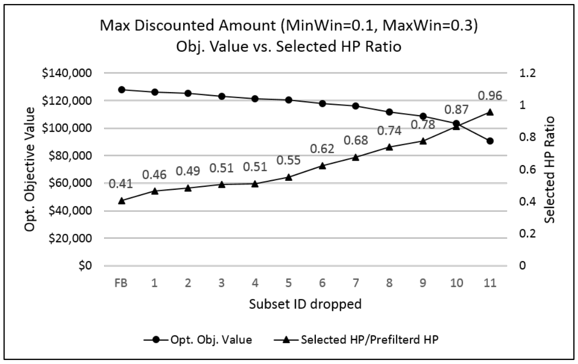


Figure 3.11: Available HPs vs. Selected HP Ratio

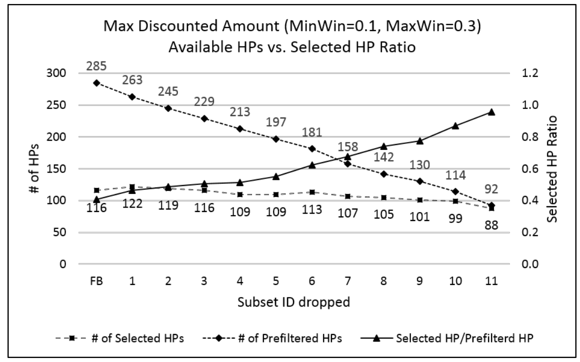


Figure 3.12: Objective Function Value vs. TPS Score

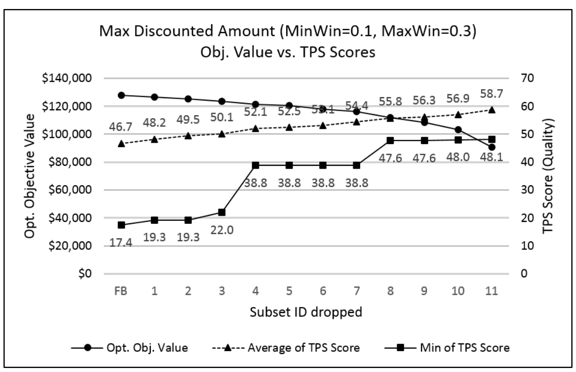
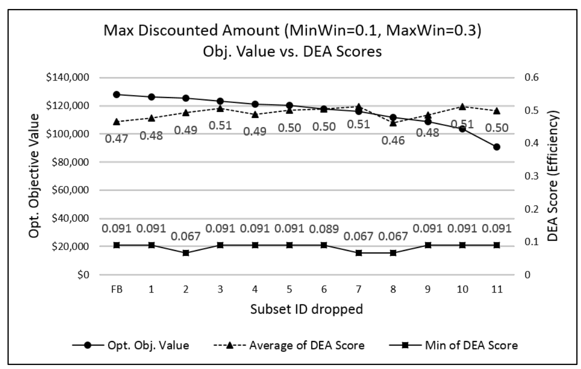


Figure 3.13: Objective Function Value vs. DEA Score

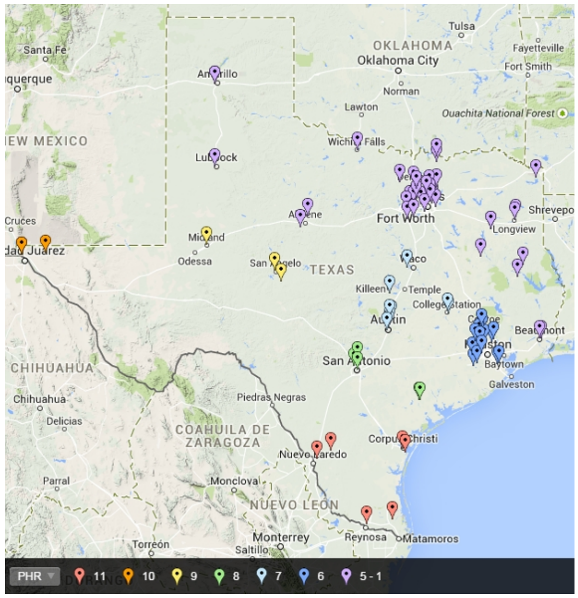


3.8 Comparison with the Existing Weighted Average Method

In this section, we compare the performance of our HP selection framework with the existing weighted average method[5]. To do this, we derive weighted average score for each HP based on three evaluation criteria, which are efficiency score (30%), quality score (35%), and discounted

[5]We mimic the evaluation criteria of ACE Demonstration which is mentioned in Section 3.2. However, it is not exactly same due to the limited accessibility of some data.

Figure 3.14: Selected Healthcare Providers in Texas (after the dropping out process until subset ID 11 in the example with $MinWin = 0.1$, $MaxWin = 0.3$)



amount (35%). Since the units for the three measures are different, we normalize them first before we derive the overall score. Finally, we derive the weighted average score for each HP by multiplying the weights to the normalized scores and summing the resulting values.

To compare with the main results elaborated in Section 3.7 (i.e., when $MinWin = 0.1$, $MaxWin = 0.3$), we determine the winners based on weighted average score under similar constraints. We define three different selected ratios (i.e., 0.1, 0.2, and 0.3) which are applied to the available HPs for each DRG code and for each PHR. We expect that as the value of selected ratio is smaller, the number of selected HP will be smaller while their average performance will be higher. The results are listed in Table 3.11. The first three columns show the results of weighted average score for each value of selected ratios, whereas the next three columns show some of the results of our HP selection framework. For example, the sub-column indicating "4" under HP Selection Framework is the result after dropping HPs assigned in subset ID 1, 2, 3, and 4 shown in Figure 3.7(b).

From Table 3.11, we can see that the minimum quality scores (i.e., *Min of TPS Score*) of selected HPs for our HP selection framework outperform the weighted average method. The minimum score after dropping subset ID 4 is always higher than 38.8, whereas the best minimum quality score for weighted average method is only 34.05 even when the selected ratio is 0.1. Thus, our selection framework effectively rules out the possibility of selecting HPs with poor quality level, while showing similar or better average of quality score compared to the weighted average method. The boxplots of TPS score and DEA score for each case in Table 3.11 is shown in Figure 3.15. The total discounted amounts of our selection framework outperform the results of weighted average method in general. For example, the total discounted amount after dropping HPs until subset ID 11 in our selection framwork is $85079.13 while selecting 88 HPs, whereas the total discounted amount of weighted average method is only $82420.96 even when the number of selected HPs is higher (i.e., 112) when selected ratio is 0.3. Overall, the HP selection framework shows stable and better results compared to the weighted average method.

Table 3.11: Weighted Average Method vs. HP Selection Framework

	Weighted Average Method			HP Selection Framework		
	0.3*	0.2*	0.1*	4**	8**	11**
# of Selected HPs	112	87	52	109	105	88
Sum of Discounted Amount	82420.96	62124.82	35878.43	120663.52	99297.58	85079.13
Average of TPS Score	53.53	54.05	55.44	52.10	55.80	58.70
Min of TPS Score	27.13	27.13	34.05	38.80	47.60	48.10
Average of DEA Score	0.57	0.62	0.70	0.49	0.46	0.50
Min of DEA Score	0.09	0.09	0.12	0.09	0.07	0.09

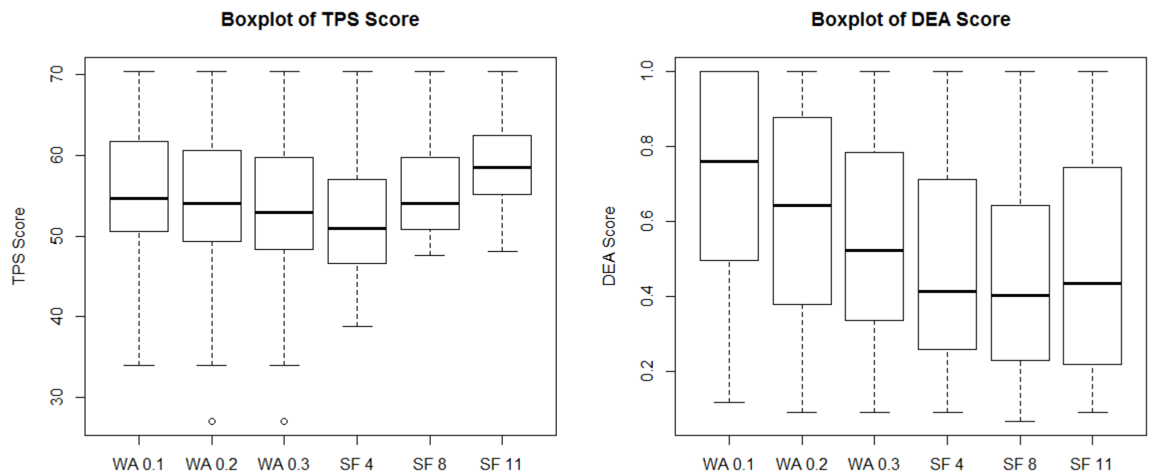
*: Selected ratio among available HPs per each DRG code and per each PHR

**: Subset ID Dropped ($MinWin = 0.1$, $MaxWin = 0.3$)

3.9 Implications and Future Research Directions

In this section, we discuss implications and suggest future directions that can help enhance the proposed selection framework. First, the performance of auctions depends on several design factors. As such, examining the impacts of the auction design issues is worthwhile future research venue that may enhance our framework. For example, we may consider the following design issues: *Which combination should be allowed in the biding process?* The auctioneer (e.g., CMS) should provide enough flexibility to form combinations of bundle payment packages in order to

Figure 3.15: Boxplots of TPS Score and DEA Score



allow the HPs (e.g., clinics, labs, and hospitals) to exploit their synergies in the bids. *How to diversify the HP base and promote competition?* The performance of a CA is influenced by the above design issues in the context of efficiency and optimality. The efficiency relates to social welfare maximization by assigning capacity to the set of HPs to achieve the most cost efficient allocation. Whereas, the optimality implies minimizing the total payments to HPs (Olivares et al., 2012).

Increasing the participation rate of the HPs is also imperative for the successful operations of bundled payments. Although the concept of bundled payment model is appealing, if only a small fraction of HPs desires to participate in the program perhaps due to misaligned incentives, the selection framework will face inefficiency. Unless the government pushes HPs by any regulations, they may decide not to join the program contrast to our assumption in this study that HPs eagerly participate in the application process. Therefore, providing analytical insights is a worthwhile future direction for incentive mechanisms that help to guarantee the high participation rate of HPs.

Bundled payment programs transfer a portion of financial responsibilities from a payer to the providers. According to the BPCI initiative evaluation report released by CMS, the change in payments is insignificant for the most of bundled episodes and the change in qualities (e.g., mortality rate and readmission rate) is also indifferent compared to FFS (CMS, 2015a, 2016a). The initiative is in its early stage so that it is hasty to make a conclusion, but certainly there is a room for improvement in provider selection and coordination methods, not only to encourage HPs to join the program but also to achieve simultaneous cost reduction and quality improvement.

To perform post hoc analysis with more convincing comparison between existing selection process and our framework, we are on the process of obtaining CMS BPCI application data. This data includes the BPCI participant list with detailed information (e.g., bundled payment care episodes, episode length, discount rate, historical average payment, risk track, episode status) and original application documents of healthcare providers that are not allowed to participate in the BPCI initiative. The original application documents includes several interesting descriptions such as care improvement plan, cost saving plan, quality measures for each care episode, and design for gain-sharing. These detailed descriptions will enable us to refine our selection framework to evaluate each applicant's improvement opportunity under BPCI along with historical performance.

3.10 Conclusion

Bundled payment models have been attempted to mitigate the deficiencies of the current fee-for-service system, and to create incentives for coordination of services needed to manage particular episodes of care. In order to achieve efficiency in delivering high quality care, determining competitive HPs becomes an important issue for government agencies such as CMS.

In this paper, we suggest a HP selection framework for bundled payment models. Our goal is to cover the demand of healthcare services in target regions by selecting required number of HPs who will participate in a bundled payment program. Since it is undesirable for applicants to focus on either cost reduction or quality of care, we develop a model which integrates both. As a result, our selection framework determines winners taking into account service quality, efficiency and cost reduction of the cares. The results of numerical analysis support the effectiveness of our selection framework compared to the existing weighted average method.

4. ADAPTIVE CAPACITY PLANNING FOR AMBULATORY SURGERY CENTERS

4.1 Introduction

Ambulatory Surgery Centers (ASCs), the healthcare facilities that provide surgical procedures exclusively on an outpatient basis, are growing trend in the U.S. These "same-day surgery" services have been made possible by technology advancement that allows a broad range of procedures to be performed safely in an outpatient setting and at a much lower cost than a hospital. For example, the Centers for Medicare and Medicaid Services (CMS) pays ASCs approximately 49% of what it would pay a hospital outpatient department (ASCA, 2011). A recent statistic shows that 5,519 ASCs in 2016 (which is a significant jump from the 1000 ASCs in 1988) in the U.S. (MedPAC, 2017) perform 23 million surgeries annually, which is equivalent to $90 billion[1]. Today, approximately 30% of Medicare beneficiaries receive care provided in the ASCs, collectively implying the importance of ASCs in the U.S. healthcare delivery system.

However, the growth of the number of ASCs has slowed in recent years[2] (MedPAC, 2017). This trend may continue because of economic conditions that are likely to discourage the establishment of new ASCs, for example, consolidation of management companies, physician employment by hospitals, and the increasing prevalence of high-deductible health plans (VMG-Health, 2017). If the supply of ASCs does not keep pace with the demand for outpatient surgery, then this care will be provided in a less convenient and more costly hospital outpatient department[3]. One possible solution is to improve the efficiency of current ASC operations by properly planning or adjusting capacity in light of the changing environment, which is the main focus of this study.

Researchers have put considerable efforts into scheduling inpatient surgeries by using mathematical programming and queueing methods in the setting of single stage and single server (Berg and Denton, 2012). However, multi-stage (i.e., pre-, intra-, and post-operative) and multi-server

[1] Oxford Outcomes ASC Impact Analysis 2010.

[2] The number of ASCs increased only 1.2% compounded annually from 5,135 in 2010 to 5,519 in 2016.

[3] Indeed, the Ambulatory Surgery Center Association analyzed that if half of the eligible surgical procedures moved from hospital outpatient departments to ASCs, Medicare would save $2.5 billion per year. Available at http://www.ascassociation.org/AdvancingSurgicalCare/whatisanasc/historyofascs.

(i.e., multiple rooms in each stage) are typical characteristics of ASCs, which make the capacity planning in ASCs more complex. Consequently, the insights from previous studies that focus on simpler problem settings may not directly carry over to ASC environments. Because of the structural complexity involved in planning ASCs, discrete event simulation has frequently been used in the literature (e.g., Tiwari and Sandberg, 2016). The simulation models are helpful to diagnose bottlenecks and to evaluate alternative scenarios, but validating the results is often challenging because simulations do not the optimal solution. Thus, a major research opportunity exists to plan capacity by using more advanced optimization approaches.

In this study, we propose a capacity planning approach for ASCs that coordinates the stages: preparation, surgical processes, and patient recovery. Hereafter, *capacity* refers to the physical space in each stage (e.g., the number of operating rooms) In general, all activities are connected closely. Furthermore, the duration of each activity is stochastic. Both dependencies and uncertainties make delivering a smooth patient flow a challenging task for which surgical centers need to manage the capacity of each activity and to schedule patients. Hence, a systematic capacity planning framework can play an influential role to improve efficiency and utilization of ASC operations. As such, we answer our main research question: *(a) Given uncertainties in patient-mix and service durations, how do we allocate capacity (the number of rooms) for an ASC that comprises three stages of patient flow, i.e., the pre-operative in a holding room (HR), the intra-operative in a operating room (OR), and the post-operative in a post-anesthesia care unit (PACU)?* To optimally design patient schedules and to plan capacity for an ASC, we analyze actual patient flow data, beginning with deterministic scenarios in both patient-mix and activity duration. This enables us to formulate exact models and to develop and evaluate heuristics designed to solve this problem. Subsequently, we extend our study to stochastic scenarios. In contrast to the extensive research that focuses on capacity allocation solely for operating rooms (e.g., Batun et al., 2011; Denton et al., 2010) or for the last two stages (e.g., Bowers, 2013; Liu et al., 2019), our study explicitly considers capacity for all three stages.

Further, we address issues related to the changing environment in ASC practice. The procedures performed in ASCs are broad in scope. Medicare currently allows for ASCs to perform more than 3,500 types of procedures, which require varying levels of time and resources in each stage, and Medicare approves new procedures for ASCs every year. Naturally, the capacity implications of adjusting the list of procedures become a relevant but challenging question for ASC practitioners. Hence, we have our second research question: *(b) What is the impact of adding/removing certain service types (or procedures) in the ASC capacity planning?* Overall, this study tackles *resource capacity planning* at the *strategic* level in the healthcare planning and control framework described by Hans et al. (2012).

In this study, we assume that the weekly patient demand is already planned by the ASC planner. The problem then is how to determine the proper balance of capacity among the stages and how to allocate such weekly demand to each weekday. Using real patient flow data, we classify patients into a fewer groups based on the duration required in each stage for their procedures. Because the OR is the most expensive resource in ASCs, we next formulate a problem that provides patients' daily schedule for given fixed OR and PACU capacities. Afterwards, we develop an algorithm that finds the optimal number of ORs and PACUs to minimize the sum of overtime cost and capacity construction cost. Lastly, we derive the minimum number of HRs that preserves the patient schedule determined in ORs and PACUs. Uncertainty in either patient-mix or durations requires additional capacity. We develop a heuristic procedure that is straightforward and easy to implement to evaluate room capacity under such uncertainties. Considering the sequential stages with multiple rooms in each stage, ASC is modeled as a *hybrid flow shop (HFS) with blocking*, i.e., a flow shop with parallel machines at each stage and with blocking constraints between any two consecutive stages (Pinedo, 2016). The *blocking* indicates the absence of storage capacity between stages, and thus, a patient must stay in an OR if all PACUs are occupied. We later show (in Theorem 2) that *blocking* in our problem setting is equivalent to a *no-wait* constraint, which is a special case of blocking. The no-wait constraint forces any two consecutive operations for a patient to be processed without any interruptions (Mascis and Pacciarelli, 2002) and describes the ideal patient flows in ASC settings better than blocking. For example, patients with anesthesia in HR should be transferred to OR right away rather than staying at HR because of lack of ORs. Similarly, patients after surgery should be moved to PACU without being delayed at OR. Unlike the majority of literature that assumes given fixed capacity, our framework is the first three-stage hybrid flow shop in the context of capacity planning.

The remainder of this paper is organized as follows. In Section 4.2, a literature review on ASC planning and scheduling is presented. In Section 4.3, we describe our problem, and Section 4.4 defines patient groups and patient sample paths (i.e., patient-mix for each weekday) using real patient flow data from an ASC. Next, in Section 4.5, we propose our ASC capacity planning approach. We formulate a mixed integer program (MIP) for the problem covering intra- and post-operative stages. Because the hybrid no-wait flow shop having only two stages is known to be strongly NP-hard (Sriskandarajah and Ladet, 1986), we develop an efficient heuristic procedure and compare its performance with solutions by MIP. In Section 4.6, we conduct computational experiments to test the performance of our heuristics and to provide managerial implications. Section 4.7 concludes our study and provide future research directions.

4.2 Literature Review

Extant literature examines operating room allocation, sequencing, and scheduling problems in various settings, but most papers pursue efficiency improvements under given fixed capacity. Hence, we have identified a research gap in capacity planning decisions equipped with scheduling guidelines.

In this section, we review previous work pertinent to the surgery scheduling problem with downstream PACU constraints. As there is limited research on capacity planning in this context, our review focuses more on scheduling literature. ASCs are a special case of HFS as they have multiple stages and parallel rooms in each stage. More general reviews on operating room management can be found in Cayirli and Veral (2003); Gupta (2007); Gupta and Denton (2008); Cardoen et al. (2010), and (May et al., 2011).

4.2.1 Surgery Scheduling with Downstream PACU Capacity

Compared to the considerable amount of research focusing solely on operating rooms (e.g., Denton et al., 2007; Mancilla and Storer, 2012; Mak et al., 2014), few studies consider PACU capacity in their surgery scheduling problems. The multiple-OR scheduling problem with shared downstream PACUs is significantly more challenging than scheduling patients only to multiple independent ORs. To overcome the theoretical and computational challenges, several studies use simulation. For example, Dexter et al. (2001); Marcon and Dexter (2006), and Iser et al. (2008) investigate sequencing heuristics and their impacts on the performance of ORs and PACUs.

To alleviate the computational complexities arising from the downstream PACUs, several researchers assume deterministic service times. For example, Hsu et al. (2003) formulate a multiple-OR sequencing and scheduling problem as a two-stage no-wait flow shop with deterministic service times. They propose a tabu search-based greedy heuristic to minimize the number of PACU nurses and PACU makespan. OR blocking is not allowed in their study; instead, patients are sent to the PACU immediately after surgery. Pham and Klinkert (2008) adopt the idea of flow shop planning to schedule patients having deterministic surgical durations into multiple ORs. They assign surgeries to the ORs and schedule start time for each surgery to minimize the makespan in their optimization problem. OR blocking is allowed with an assumption that recovery is not started until a patient enters the PACU. Augusto et al. (2010) also consider the problem of scheduling patients with deterministic surgical durations into multiple ORs. In their study, recovery starts in ORs if patients are blocked caused by the unavailability of PACUs. Unlike the works discussed above, the main objective of our study is to determine the capacity in each stage that minimizes the sum of overtime cost and capacity construction cost, rather than the makespan. In capacity planning perspectives, having enough PACUs can simply resolve the OR blocking problem. In other words, considering OR blocking that allows patients to recover in ORs is important only when PACU capacity is fixed and thus cannot be adjusted. Hence, in our study, we rely on no-wait assumption between ASC stages and consider expanding PACU capacity if OR blocking is expected.

More recently, researchers consider the randomnesses of medical service times and the downstream PACU capacity together, which are closely related to our research. Lee and Yih (2012) investigate the problem of scheduling a pool of surgeries with lognormally distributed durations into multiple ORs with a limited and fixed number of downstream PACUs. The surgery-to-OR assignment and the scheduled start times of the surgeries in each OR are obtained using a genetic algorithm (GA) to minimize the total cost associated with patient waiting, OR blocking, OR idle time, and OR overtime. Lee and Yih (2014) determine scheduled start times for surgeries in multiple ORs constrained by limited PACU capacity when the sequence of surgeries in each OR is given. Surgical service times are assumed to be fuzzy triangular numbers. Bai (2017) proposes both sequencing and scheduling of surgeries in multiple ORs with limited PACU capacity. They propose a two-stage solution method to minimize the expected cost of patient waiting time, surgeon idle time, OR blocking time, OR overtime, and PACU overtime. A mixed integer linear programming sample average approximation model is formulated and solved by Lagrangian relaxation in the first stage to obtain the sequence of surgeries in each OR; then a sample-gradient-based algorithm developed in Bai et al. (2017) is used in the second stage to schedule the start times of surgeries. In contrast to the previous works cited above that focus on efficient patient sequencing and scheduling under given capacity, our study focuses on capacity planning for the multi-stage nature of ASCs. In Table 4.1, we compare our problem settings with closely related previous studies. Rather than making scheduling decisions based on available capacity, our study proposes how ASC planners make investments to construct/renovate/adjust the system capacity and thereby improve the ASC operations.

4.2.2 Planning Capacity and Scheduling Patients in ASCs

Unlike scheduling papers that provide mathematical models, previous studies on capacity planning mostly rely on simulations or heuristics. For example, Tiwari and Sandberg (2016) examine whether the current capacity in the pre- and post-operative stages of an ASC is enough to cover a given surgical demand with the assumption that ORs are always occupied during regular daily hours. Motivated from this study that investigates the appropriate level of ASC capacity, we use mathematical programming to incorporate patient scheduling, which provides a stepping stone to

Table 4.1: Related Literature in Different Settings

Problem Settings	Our Paper	Bai (2017)	Bai et al. (2017)	Lee and Yih (2014)
Considered Resources				
Upstream HR	Yes	No	No	No
OR	Yes	Yes	Yes	Yes
Downstream PACU	Yes	Yes	Yes	Yes
Focus				
Surgery-to-OR Assignment	Yes	No	No	No
Sequencing	Yes	Yes	No (Given Sequence)	No (Given Sequence)
Scheduling	Yes	Yes	Yes	Yes
Capacity Planning	Yes	Fixed	Fixed	Fixed
Patient-Mix				
Patient Grouping	Yes	No	No	No
Cost Structure				
Capacity Construction Cost	Yes	No	No	No
Patient Waiting Time	No	Yes	Yes	Yes (2nd Stage)
OR Blocking Time	No	Yes	Yes	Yes (1st Stage)
OR Idle Time	Potentially Yes	Yes	Yes	Yes (2nd Stage)
OR Overtime	Yes	Yes	Yes	No
PACU Overtime	Yes	Yes	Yes	No
Notable Constraint	No-wait between Stages	OR Blocking Allowed	OR Blocking Allowed	OR Blocking Allowed
Structural Properties	Yes	No	Yes	No
Solution Approach	Heuristics and Iterative Algorithm	Subgradient method	SAA-gradient descent algorithm	Genetic Algorithm, Heuristics

make the capacity decision.

Metaheuristics are also often used to evaluate capacity decision. For example, Gul et al. (2011) study a multiple-OR planning problem with pre-operative and post-operative resource constraints. Surgeries are first assigned to ORs and then sequenced and scheduled in ORs. Their bi-criteria genetic algorithm method is effective in surgery re-assignment but does not show significant improvements over the shortest-case-first heuristic if surgery-to-OR assignments are fixed.

Cayirli et al. (2006) demonstrated that the performance of an appointment schedule is very sensitive to patient characteristics and behavior such as walk-ins (add-ons or patients that need to be scheduled the same day), no-shows, and punctuality. Although the authorsâĂŹ analysis was specific to a clinical office setting (primary care), the role of patient characteristics and behavior in scheduling decisions is also informative to ASCs settings.

There are several other studies that shed light on one aspect of the capacity allocation problem but ignore one or more important features. For example, using stochastic programming, Min and Yih (2010) generate an optimal surgery schedule for elective patients assuming uncertainty in surgery durations and thereby the availability of downstream surgical intensive care units (ICUs). To reduce problem complexity, the authors assume that surgery and downstream resources can be scheduled independently. In our ASC problem, however, we assume no-wait between stages. In other words, the scheduling of OR directly impacts the PACUs' schedules. Price et al. (2011) apply integer programming and simulation to develop improved surgical scheduling assignments to ultimately reduce the boarding of patients overnight in the PACU resulting from the lack of ICU resources. The patient flow in their study, however, begins at OR, not upstream at HR, which is one of our contributions.

4.2.3 Hybrid Flow Shop with No-Wait Constraints

As discussed earlier, the ASC patient flow structure that comprises three sequential stages with multiple rooms in each stage can be modeled as HFS. Given its importance and complexity, the HFS problem has been extensively studied, mainly in the manufacturing context. Ruiz and Vázquez-Rodríguez (2010) and Ribas et al. (2010) present comprehensive literature reviews on exact and heuristic methods that have been proposed to solve HFS problems under various problem settings. For example, two- and three-stage problems are approached by Koulamas and Kyparisis (2000) with the aid of simple methods with known performance bounds. Interestingly, however, none of the previous studies proposes an approach to decide the number of servers/machines in each stage, which is our focus.

In the literature, HSF with no-wait constraint is closely related to our study. No-wait implies that the operations of a job have to be processed from the beginning to the end without any interruption on machines and without any waiting in between the machines. For example, Guirchoun et al. (2005) model a computer system composed of a single server and two identical parallel machines as a two-stage HFS with the no-wait constraint between the two stages.

The standard form of HFS relies on the following assumptions: all jobs and machines are available at time zero, the machines at a given stage are identical, any machine can process only one operation at a time, and any job can be processed by only one machine at a time; setup times are included in the processing time, preemption is not allowed, the capacity of buffers between stages is unlimited, and problem data is deterministic and known in advance (Ruiz and Vázquez-Rodríguez, 2010). Reflecting ASC practices, our study relaxes the last two of the common assumptions. We do not allow patient waits between any two stages (i.e., there is no buffer). Although we initially build our model for deterministic cases, we also address the capacity implications when the patient-mix information is stochastic following a normal distribution and the patient duration at each stage is stochastic following a log-normal distribution.

4.3 Problem Description

The main objective of this study is to provide a framework to plan capacity for ASCs in which the patient demand over weekdays is set up by top management and thus given. As all patients visiting ASCs are elective, ASC planners generally have control over the planned patient demand and hence such assumption on exogenous patient demand is acceptable. The problem is then to determine the capacity, i.e., the number of rooms, in each stage that efficiently covers possible daily patient demand and to coordinate the three stages of patient visits: (i) pre-operative (preparation),

Table 4.2: Classification of papers related to our study

ASC or Outpatient Clinical Center		Hybrid Flow Shop, No-wait	
Long-term Planning	Short-term Scheduling	2-stage	3-stage and more
Liu et al. (2010), Patrick et al. (2008), Dexter and Traub (2002), Price et al. (2011), Adan et al. (2009), Gerchak et al. (1996), Muthuraman and Lawley (2008), Lee et al. (2018), Berg et al. (2011), Hsu et al. (2003)	Berg et al. (2010), Robinson and Chen (2003), Gul et al. (2011), Denton et al. (2007), Cayirli et al. (2006), Min and Yih (2010)	Guirchoun et al. (2005) (Exact), Wang et al. (2015), Yang (2010), Wang and Liu (2013), Chang et al. (2004), Wang et al. (2005), García and Lozano (2005), Xie et al. (2004) (Heuristics)	Gicquel et al. (2012), Liu and Karimi (2008) (MIP), Thornton and Hunsucker (2004) (Heuristics)
Our Paper	Our Paper		Our Paper

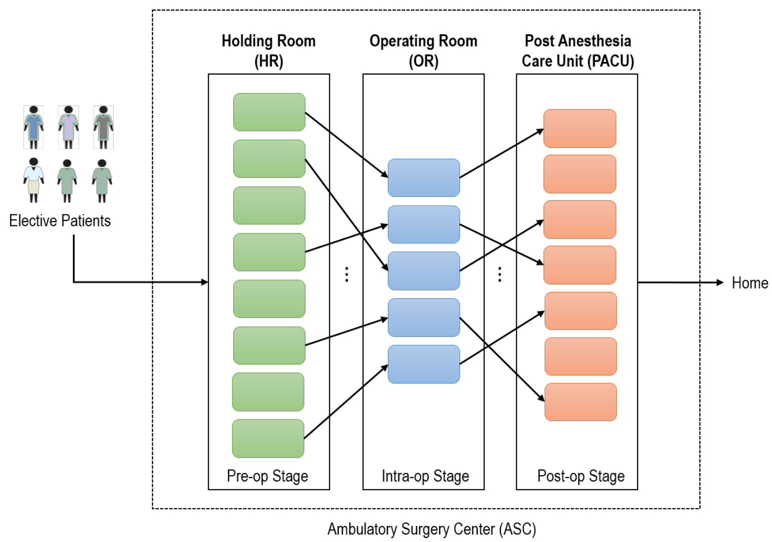
(ii) intra-operative (surgical procedures), and (iii) post-operative (recovery) stages. We use the term *room* for ease of reading, and it includes all the equipment and the bed needed for the room. The objective is to minimize the total cost incurred in satisfying the daily patient demand, where the total cost is defined as the sum of overtime cost and amortized construction cost for the three stages. The overtime cost can occur in any stage when patients are served after regular hours and the construction cost is essentially the fixed setup cost of a room. Clearly, trade-offs exist between the two cost measures. Note that the daily patient-mix and the activity duration in each stage have inherent variation. Such uncertainty in patient-mix and activity duration makes achieving smooth and efficient patient flow a challenging task. Analyses must consider all of these factors in devising capacity plan for ASCs (Green, 2002). ASCs may improve cost efficiency by properly allocating capacity for each stage and by efficiently scheduling patients.

In the *pre-operative* stage, a patient requires approximately 2 hours for preparation (e.g., providing anesthesia to a patient, etc.) at a Holding Room (HR). The *intra-operative* stage is performed in an Operating Room (OR) with an average surgery duration of 1.5 hours. To attenuate the complexity involved in each patient type (i.e., service type), we adopt a patient categorization strategy based on the surgery duration, similar to Price et al. (2011) and Adan et al. (2009), after analyzing the distribution of the durations across patient types. After surgery, the patients are transferred to a recovery room, i.e., a PACU. In this *post-operative* stage, on average, a patient takes 2 to 3 hours for recovery, while some patients stay in PACU and recover over night.

In HR, unless a nurse works more than 10 hours in a day, which rarely occurs, there is no overtime. Patients are assigned to HR starting at 5 am. In OR, regular hours span from 7 am to 5 pm (i.e., 10 hours). Thus, we consider any operating procedures after 5 pm as overtime. Similarly, we follow regular hours from 7 am to 5 pm for PACUs. If overnight-stay patients exist, we consider the extra cost per patient occupying PACUs (e.g., wage of PACU nurses) as an overtime cost.

Figure 4.1 summarizes patient flows at the ASC. Figure 4.2 provides example major time stamps at the ASC, e.g., HR enter, OR enter, and OR exit times (see Appendix C.2 for the detailed list of events). Based on these time stamps, we calculate overtime in each stage.

Figure 4.1: Patient Flow at the Ambulatory Surgery Center



By analyzing real-world patient flow data from an ASC, we propose a generic capacity planning framework built upon scheduling theory and applicable to any ASCs once planned patient demand is determined. Beginning with deterministic data on patient groups and their durations at each stage, we gradually relax these assumptions to provide insights on ASC capacity planning if the composition of patient groups changes over time with uncertain durations, which are common challenges for the ASC planners.

Figure 4.2: Major Time Stamps at the Ambulatory Surgery Center

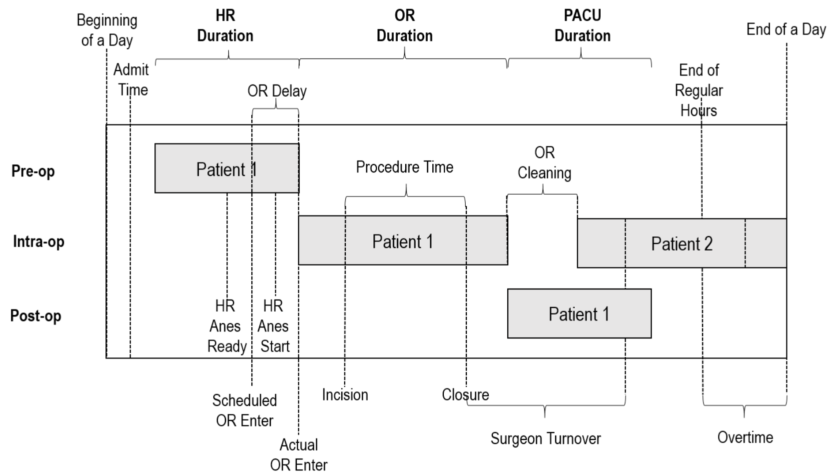
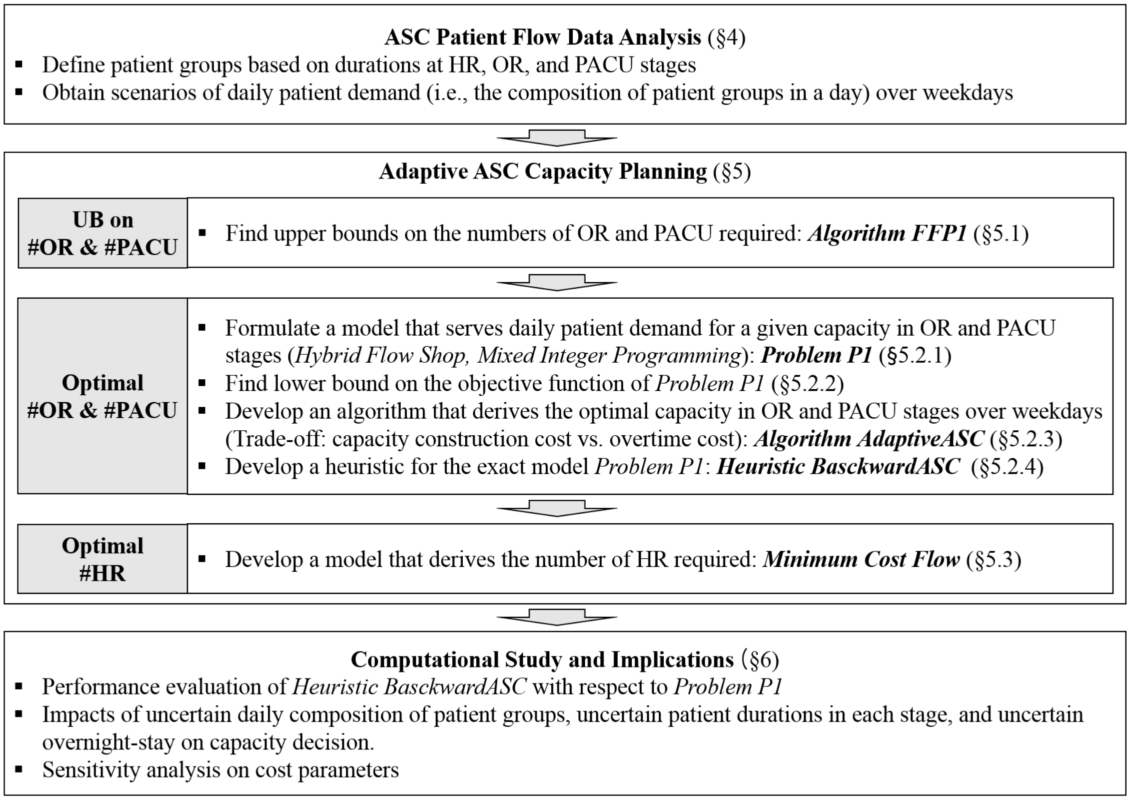


Figure 4.3 summarizes the sequence of our adaptive capacity planning framework. First, using historical patient flow data, we classify the patients into a small number of groups and obtain scenarios of daily patient demand (namely, patient sample paths) (Section 4.4). This process is designed to alleviate complexity in patient-mix to assist planners in defining their patient demand. Second, assuming deterministic patient sample paths and their durations at each stage, we develop an exact model, Problem *P1*, and an algorithm, *AdaptiveASC*, that derive the most efficient capacity in each stage of an ASC (Section 4.5). The key trade-off in our decision making is between the capacity construction cost and the overtime cost that may occur if capacity is not enough. Third, to overcome the computational challenges of the exact approach, we develop heuristics designed to achieve the same objective. In brief, our adaptive capacity planning approach evaluates performance of an ASC with given capacity, and iteratively adjusts capacity in each of the ASC stages as long as it is cost-efficient to do so. Lastly, we conduct computational experiments to address scenarios with stochastic patient sample paths and stochastic durations to provide managerial insights for ASC planners.

Figure 4.3: Sequence of the ASC Capacity Planning Framework



4.4 Defining Patient Groups and Patient Sample Paths

The composition of daily patient-mix affects the capacity planning of ASCs because patients with varying conditions require different amount of resources and time for treatment. We thus begin with analyzing the detailed time stamp data[4], enabling us to calculate the duration for each stage and for each patient. The ASC operates weekdays (a week consisting of 5 days) and the average number of outpatients per day is approximately 30 to 35 patients. In general, ASCs are segmented into either single-specialty ASCs or multi-specialty ASCs. In this study, we focus on multi-specialty ASCs as the multi-specialty ASCs are a more lucrative segment (FMI, 2017) and the findings from multi-specialty ASCs can carry over to the relatively less complex single-specialty ASCs. The ASC that provides patient flow data to us has 23 HRs, 11 ORs, and 12

[4]We obtained the data from an ASC located in the southeast of the U.S.

PACUs[5], and provides services for a broad spectrum of patients. Our data, spanning from December 2016 to February 2017, contains 1961 patients who fall into 13 service types including otolaryngology, ophthalmology, general oncology surgery, orthopedic, and plastic surgery. Specifically, our data contains patients with 523 different Current Procedural Terminology (CPT) codes (See Appendix C.3).

We first define *patient groups* based on the duration of patient stay in each stage. Subsequently, we analyze the daily patient demand, defined as a combination of patients each of which falls into one of the patient groups, namely *patient sample path*, to derive optimal capacity. In our data, we observe varying daily patient sample paths (i.e., scenarios) representing uncertain daily patient demand and its associated durations.

Specifically, we first analyze the distributions of patient duration at HR (pre-op), OR (intra-op), and PACU (post-op) stages. We assume that the transportation time from HR to OR (from OR to PACU, respectively) are included in HR duration (PACU duration, respectively). Previous studies (e.g., Gul et al., 2011; Akcali et al., 2006) collectively provide evidence that the surgery time depends on the surgical procedure, i.e., the complicated surgery may require longer operating time. In contrast, the relationship between the durations between surgical procedure and pre-op/post-op varies depending on problem settings. For example, Akcali et al. (2006), examining resource allocation at the operating room and the subsequent ICU for cardiohepatic surgeries, categorized patients into eight groups based on the duration of patient stay at those two stages. The pre-op stage is not discussed in this study. In a later work, Gul et al. (2011) observe that pre-op and post-op durations are similar regardless of surgical procedures and thus they assume the same triangular distributions for both pre- and post-op durations in the simulation study. In our ASC patient flow data, we observe a low correlation between durations at HR and OR (i.e., $r = 0.08$), but a moderate correlation between the duration of OR and PACU (i.e., $r = 0.40$). Besides, the total duration is highly correlated with the durations at OR (i.e., $r = 0.58$) and PACU (i.e., $r = 0.96$), but not with the duration at HR (i.e., $r = 0.25$) (See Table 4.3). The HR durations

[5] As mentioned in Tiwari and Sandberg (2016), however, up to 14 out of 23 HRs can be flexibly used as PACUs for overnight-stay patients.

Table 4.3: Correlation Table for Duration in Each Stage

		(1)	(2)	(3)	(4)	(5)	(6)	(7)
(1)	HR Duration	1	**0.08**	0.05	-0.01	0.04	0.04	**0.25**
(2)	OR Duration		1	0.36	-0.10	0.38	**0.40**	**0.58**
(3)	Phase 1 Recovery Duration			1	-0.20	0.24	0.34	0.38
(4)	Phase 2 Recovery Duration				1	-0.28	-0.17	-0.17
(5)	Phase 3 Recovery Duration					1	0.99	0.94
(6)	PACU Duration						1	**0.96**
(7)	Total Duration							1

Notes: $p < 0.05$ if $|r| > 0.04$. PACU duration is sum of phase 1, 2, and 3 durations. Phase 3 durations may contain overnight-stay but not necessarily.

of our patient data seem to be independently distributed from OR durations (e.g., Gul et al., 2011) but the PACU durations are closely related to the OR durations (e.g., Akcali et al., 2006). We thus use the OR and PACU durations together to classify patients into a small number of groups, while estimating a single distribution for HR durations of all patients in our data.

Classifying patients into a reasonably small number of groups not only reduces the complexity of the capacity planning problem but also improves its practicality. For example, ASC managers may form a patient sample path (e.g., reflecting surgeons' availability/preference over weekdays) from a pool of patients waiting for surgery operations. In our study, we classify patients into short/moderate/long OR duration and short/moderate/overnight PACU duration. By assigning the diagnoses of patients to the duration-based groups, ASC managers can easily classify any new patients into one of the patient groups. We use the patient groups as a basic building block forming a patient sample path, i.e., the composition of patients, and ultimately to plan capacity of an ASC.

To classify patients, we first conduct k-means clustering for each service type. To properly allocate resources, we further consider whether each patient stays overnight or not. We thus conduct a cluster analysis based on average OR duration and average PACU duration for each surgical procedure (i.e., CPT code) that falls within a focal service type. We further divide each CPT code into two subgroups: with and without overnight-stay because overnight-stays require much longer durations at PACUs[6]. As a result, we obtain seven patient groups as listed in Table 4.4. For example, the patient group *ShortOR-ShortPACU* comprises patients with relatively shorter OR durations (with on average 68.4 minutes) and PACU durations (with on average 66.3 minutes) than other patient groups. Similar to previous studies, either lognormal, gamma or Weibull distribution fits well for the duration at any stages. The fitted duration distributions for each patient group are provided in Appendix C.3.

Table 4.4: Defined Patient Groups (in min)

Patient Group	Duration at OR	Duration at PACU	# Serv -Clus[a]	Obs.	Avg OR	Med[b] OR	Avg Ph[c]1&2	Avg Ph3	Avg PACU	Med PACU	Avg Total
1	Short	Short	5	727	68.4	63.0	63.6	2.6	66.3	58.0	265.3
2	Short	Moderate	4	333	77.9	71.0	87.4	11.7	99.1	82.0	301.8
3	Moderate	Short	3	154	137.7	134.0	88.0	11.7	99.7	85.0	365.9
4	Moderate	Moderate	4	512	145.3	136.0	113.5	11.4	124.9	110.0	419.1
5	Moderate	Overnight	3	33	166.3	164.0	91.8	1063.8	1155.6	1175.0	1461.7
6	Long	Moderate	5	74	298.6	278.5	104.9	52.6	157.5	142.0	589.3
7	Long	Overnight	4	127	225.3	214.0	76.4	1124.9	1201.3	1209.0	1579.7

Notes. [a]: # Service Type Clusters. See Table C.3; [b]: Median; [c]: Phase;

Based on the patient groups that we defined above, we next summarize the patient sample paths across weekdays as listed in Table 4.5. The patient sample path represents the composition of daily patients from the seven patient groups. On Monday, for example, on average 11.45 *ShortOR-ShortPACU* patients and 8 *ShortOR-ModeratePACU* patients have visited the ASC. Note that, however, these patient sample paths are obtained mainly based on ongoing daily scheduling approach at the ASC. At the moment, we use the patient sample paths obtained from our patient flow data (as listed in Table 4.5) to feed our capacity planning model. In our computational study, we also test our model on various patient sample paths. As such, we suggest guidelines that ASC

[6]We follow this patient clustering approach based on the "average" duration of each CPT code rather than individual patient-level duration to avoid possibility that patients with the same CPT code being classified into different clusters. To decide reasonable k, which is a tricky part of using k-means clustering, we test several options recommended by the literature, and then choose the final k after analyzing plots and characteristics of each cluster (See Appendix C.3 for details).

Table 4.5: ASC Patient Sample Path by Weekday (# of patients)

	Patient Group	Monday	Tuesday	Wedesday	Thursday	Friday
1	ShortOR-ShortPACU	11.45 (3.33)	15.08 (3.52)	12.17 (4.06)	7.92 (2.75)	12.58 (4.78)
2	ShortOR-ModeratePACU	8.00 (2.83)	3.62 (2.33)	5.50 (2.47)	6.15 (2.70)	4.33 (2.74)
3	ModerateOR-ShortPACU	4.00 (2.05)	3.08 (2.10)	1.92 (1.08)	1.69 (1.84)	1.42 (1.08)
4	ModerateOR-ModeratePACU	3.64 (2.46)	8.23 (4.04)	10.75 (3.49)	10.15 (2.70)	8.67 (1.78)
5	ModerateOR-OvernightPACU	0.73 (0.65)	0.23 (0.44)	0.33 (0.65)	0.77 (0.73)	0.67 (0.78)
6	LongOR-ModeratePACU	2.09 (0.70)	1.00 (0.82)	0.42 (0.51)	1.77 (1.17)	0.83 (1.03)
7	LongOR-OvernightPACU	3.18 (1.40)	1.85 (1.77)	2.08 (1.38)	2.23 (1.24)	1.17 (1.03)
	# Complete Weeks in Our Data	11	13	12	13	12

Notes: For example, on Monday, there are on average 11.45 patients classified into Patient Group 1 with standard deviation of 3.33, where as on Thursday, there are 7.92 Group 1 patients on average with standard deviation of 2.75.

Table 4.6: Duration at each Stage including Turnover Time (in hours) and Patient Sample Path on Monday

Duration + Turnover Time	Patient Group						
(in hour)	1	2	3	4	5	6	7
OR	1.5	2.0	3.0	3.0	3.5	5.5	4.5
PACU	1.5	2.0	2.0	2.5	19.5	1.0	19.0
Non-overnight PACU	1.5	2.0	2.0	2.5	2.0	1.0	1.5
Overnight PACU	-	-	-	-	17.5	-	17.5
Patient Sample Path (# of patients)	11	8	4	4	1	2	3

managers can easily incorporate to define a set of patients to be served in a day.

For ease of exposition, we display the duration times in hours and round them off to the bins with size of 30 minutes as in Table 4.6. The table provides durations including turnover times, i.e., 30 min in OR and 15 min in PACU durations (e.g., OR duration + Turnover time in OR for a *ShortOR-ShortPACU* patient is $68.4+30 = 98.4$min ≈ 1.5hr). The last row of Table 4.6 represents an Monday patient sample path[7]. We later use this table to feed and evaluate our heuristics.

The demand for surgeries in outpatient settings continues to grow. Hence, the importance

[7]We use rounded mean number of patients for each patient group (e.g., 11 *ShortOR-ShortPACU* patients) as the average number of patients.

of ASC capacity planning that accommodates timely access to services while utilizing resource efficiently also continues to grow. We tackle this issue by starting with the deterministic patient sample path for each weekday and deterministic duration at each stage. As can be seen later, our capacity planning approach can be adapted easily to the environment with stochastic patient sample paths and durations to provide capacity implications of such uncertainties.

4.5 Adaptive Capacity Planning Framework

In this section, we discuss our capacity planning approach. Because of the three-stage structure of an ASC, which is well-known for high complexity in HFS literature, there is a limited chance of using exact methods. More importantly, the capacity, which is our decision variable, is a given parameter in HFS formulations (Ruiz and Vázquez-Rodríguez, 2010). As a result, how to decide the right capacity is rarely discussed in the literature. Our approach to deal with this challenge starts from splitting the three stages in to two parts: (1) derive the capacity in OR and PACU stages; (2) derive the capacity in HR stage. Since we focus on the later two stages first, we call this approach backward capacity planning[8].

In Part (1), we first obtain the upper bound on the numbers of ORs and PACUs (Section 4.5.1). Next, we propose a capacity planning approach to find optimal numbers of ORs and PACUs (Section 4.5.2). Specifically, we formulate a mixed integer program (MIP), Problem *P1*, that generates the schedule of planned daily patient demand for given numbers of OR and PACU (Section 4.5.2.1). Afterwards, we develop an algorithm *AdaptiveASC* that iteratively evaluates capacity to provide the most cost-efficient combination of the numbers of ORs and PACUs under the trade-off between capacity construction cost and overtime cost (Section 4.5.2.2). Because of the computational complexity in solving Problem *P1* in the algorithm, we derive lower bounds on the objective function of *P1* (Section 4.5.2.3) and develop a heuristic *BackwardASC* to replace the exact model Problem *P1* (Section 4.5.2.4).

Afterwards, in Part (2), using a minimum cost flow model, we derive the optimal number of

[8] One might wonder what if we plan capacity in forward approach, i.e., we focus on the HR and OR stages first and determine the PACU stage later. Since the OR and PACU stages are relatively more costly than the HR stage, we find evidence that the backward approach outperforms the forward approach. See Appendix C.8 for details

HR required to preserve the patient schedules obtained in Part (1) (Section 4.5.3). Ultimately, our approach not only derives a cost-efficient ASC capacity but also provides patient scheduling guidelines under such capacity, which are important strategic and operational issues for ASC planners.

4.5.1 Upper Bounds on the Number of ORs and PACUs

Algorithm *FFP1* derives upper bounds on the numbers of ORs and PACUs. We modify the "First Fit" (FF) strategy used for the well-known bin packing problem to obtain a feasible assignment of patients into the daily regular hours of our problem setting. By running Algorithm *FFP1*, ASC managers may find the largest capacity for OR and PACU stages to serve the planned daily patient demand.

Algorithm First Fit for Problem *P1* (*FFP1*)

Input: The set of I patients and a bin consists of a pair of OR and PACU. Initialize the bin size to T. The bin is a two-machine, no-wait flowshop, $F_2|$no-wait, $m_1 = 1,\ m_2 = 1|C_{max}$.

Step 0: We need to allocate patients into each bin such that patients are scheduled in no-wait manner (between OR and PACU) and their total completion time is no more than T. Initially, open only one bin.

Step 1: Number the given set of patients in random order, $i \in \mathcal{I} = \{1, 2, \cdots, I\}$. Patient i's duration times in OR and PACU are p_{i1} and p_{i2}, respectively. If patient i requires overnight-stay in PACU, then set $p_{i2} = 0$. Let n_o be the number of patients requiring overnight-stay in PACU. Let $i = 1$.

Step 2: Begin to schedule patient i in the first bin opened at the earliest available time. If patient i violates the threshold completion time T for either OR or PACU, then schedule the patient in the next opened bin (at the earliest available time) until patient i does not violate the threshold completion time T of that bin. If Patient i does not fit into any open bin, then open a new bin and schedule patient i in the new bin.

Step 3: Set $i = i + 1$. If $i \leq I$ go to Step 2.

Step 4: Register the number of bins opened during the process of scheduling all patients $i \in \mathcal{I}$ as Λ.

Output: The number of ORs is Λ, and the number of PACUs is $\Lambda + n_o$. Stop.

We initialize the number of ORs as $R_1 = \Lambda$ and PACUs as $R_2 = \Lambda + n_o$. We discuss the details in Section 4.5.2.2, but such initialization of ORs and PACUs is useful to efficiently run Algorithm *AdaptiveASC*, which finds the optimal number of ORs and PACUs. A numerical illustration of Algorithm *FFP1* is given in Appendix C.5.

4.5.2 Planning Capacity of OR and PACU

In this section, we elaborate our exact model Problem *P1* for given capacity in OR and PACU, derive lower bounds on the objective function, then propose a polynomial time algorithm *AdaptiveASC* that iteratively updates the capacity in OR and PACU stages until it finds an optimal solution. Finally, we develop our heuristic *BackwardASC* to replace the computationally challenging exact model *P1*.

4.5.2.1 Mixed Integer Program Problem P1

Although our ultimate goal is to decide capacity in each stage, the complex interactions across multiple stages in an ASC require significant coordination and thus convey the challenges in formulating a single model for simultaneously scheduling patients and deciding the number of rooms in each stage.

We thus solve the two-stage HFS problem (Problem *P1*) with the predetermined capacity in ORs and PACUs. Building upon this formulation that enables us to consider the trade-off between the capacity construction cost and the overtime cost of the two stages, we later propose an algorithm (Algorithm *AdaptiveASC*) that iteratively adjust the capacity in each stage to find the capacity that minimizes the sum of capacity construction cost and overtime cost.

We assume that patients are punctual, hence, there are no delays because of the late arrival of patients. All the ASC operations are elective in nature, so there is no capacity assigned to emergency patients[9]. The time horizon is discretized into time periods $\{0, \cdots, K\}$, where K is an upper bound on the last patient's completion time (i.e., makespan). No idle time is allowed between OR and PACU for each patient. Overtime occurs if completion time in any stage exceeds the regular close of business T. The notations used in Problem *P1* are listed in Table 4.7. For

[9]ASCs may have high priority add-on cases that need to be fit into the current day's schedule, but such cases rarely happen. Thus, for ease of exposition, we ignore this functionality in our study.

simplicity, we assume the same unit overtime cost C_s^o per nurse for one hour across available rooms in stage s[10]. The cost parameters are assumed to follow $C_1^e > C_2^e > C_1^o > C_2^o$, which generally holds in practice. Note that we focus on minimizing the overtime cost in ORs and PACUs together with the amortized construction cost of them, but Problem *P1* can be easily modified to include idle time cost, i.e., $\sum_{s \in \mathcal{S}} \sum_{r \in \mathcal{R}_s} \left[C_s^d (T - \sum_{t=0}^{T-1} \sum_{i \in \mathcal{I}} y_{isr}^t) \right]$, in the current objective funtion (4.1), where C_s^d is the unit idle time cost in stage s. We later show in Theorem 3 that problems with and without idle time cost in the objective function are equivalent if certain conditions hold.

Table 4.7: Notations for the time-based formulation of *P1*

Indices:	
i	index for a patient, $i \in \mathcal{I} = \{1, \cdots, I\}$.
s	index for a stage, $s \in \mathcal{S} = \{1, 2\}$ where 1=OR, 2=PACU stage, respectively.
r	index for a room, $r \in \mathcal{R}_s = \{1, \cdots, R_s\}$.
t	index for a time, $t \in \mathcal{K} = \{0, \cdots, K\}$
Parameters:	
R_s	Number of rooms in stage s.
p_{is}	Duration of patient i at stage s.
C_s^e	Amortized daily capacity construction cost for a room at stage s.
C_s^o	Unit cost of overtime in stage s.
K	Upper bound of completion time for all patients (i.e., makespan), $K \geq \sum_{i \in \mathcal{I}} \sum_{s \in \mathcal{S}} p_{is}$.
T	Length of regular hours. E.g., 10 hours (7am to 5pm).
Decision Variables:	
x_{isr}^t	If $x_{isr}^t = 1$, patient i starts processing on room r in stage s at time t, otherwise 0.
y_{isr}^t	If $y_{isr}^t = 1$, patient i uses room r in stage s at time t, otherwise 0.
z_s^t	The number of occupied rooms at time slot t in stage s. Note that $z_{st} \leq R_s$.
f_{is}	Completion time of patient i in stage s.
g_{sr}	Completion time of room r in stage s.
h_{sr}	Overtime of room r in stage s.
ω_{sr}^t	Binary variable in if-else statement.

[10]Our method can, however, be easily modified to handle distinct unit costs across rooms even in a same stage.

Problem *P1*:

$$\min_{x_{isr}^t, y_{isr}^t, z_s^t, f_{is}} \quad \underbrace{\sum_{s\in\mathcal{S}}\sum_{r\in\mathcal{R}_s} C_s^o h_{sr}}_{\text{Overtime Cost}} + \underbrace{\sum_{s\in\mathcal{S}} C_s^e R_s}_{\text{Amortized Capacity Construction Cost}} \tag{4.1}$$

subject to

$$\sum_{r\in\mathcal{R}_s}\sum_{t=0}^{K-1} x_{isr}^t = 1,\ \forall i\in\mathcal{I},\ \forall s\in\mathcal{S}, \tag{4.2}$$

$$\sum_{i\in\mathcal{I}}\sum_{t'=\max\{0,t-p_{is}\}}^{t-1} x_{isr}^{t'} \leq 1,\ \forall s\in\mathcal{S},\ \forall r\in\mathcal{R}_s,\ \forall t\in\{1,\cdots,K\}, \tag{4.3}$$

$$\sum_{t'=t}^{\min\{K-1,t+p_{is}-1\}} y_{isr}^{t'} \geq p_{is}x_{isr}^t,\ \forall i\in\mathcal{I},\ ,\forall s\in\mathcal{S},\ \forall r\in\mathcal{R}_s,\ \forall t\in\{0,\cdots,K-1\}, \tag{4.4}$$

$$\sum_{i\in\mathcal{I}}\sum_{r\in\mathcal{R}_s} y_{isr}^t = z_s^t,\ \forall s\in\mathcal{S},\ \forall t\in\{0,\cdots,K-1\}, \tag{4.5}$$

$$\sum_{t=0}^{K-1} z_s^t = \sum_{i\in\mathcal{I}} p_{is},\ \forall s\in\mathcal{S}, \tag{4.6}$$

$$\sum_{r\in\mathcal{R}_s}\sum_{t=0}^{K-1}(t+p_{is})x_{isr}^t \leq f_{is},\ \forall i\in\mathcal{I},\ \forall s\in\mathcal{S}, \tag{4.7}$$

$$\sum_{r\in\mathcal{R}_1}\sum_{t=0}^{K-1} t x_{i2b}^t \geq f_{i1},\ \forall i\in\mathcal{I}, \tag{4.8}$$

$$f_{i2}-(f_{i1}+p_{i2}) \leq 0,\ \forall i\in\mathcal{I}, \tag{4.9}$$

$$(t+1-g_{sr}) \leq K\omega_{sr}^t,\ \forall s\in\mathcal{S},\ \forall r\in\mathcal{R}_s,\ \forall t\in\{0,\cdots,K-1\}, \tag{4.10}$$

$$\sum_{i\in\mathcal{I}} y_{isr}^t \leq K(1-\omega_{sr}^t),\ \forall s\in\mathcal{S},\ \forall r\in\mathcal{R}_s,\ \forall t\in\{0,\cdots,K-1\}, \tag{4.11}$$

$$g_{sr}-T \leq h_{sr},\ \forall s\in\mathcal{S},\ \forall r\in\mathcal{R}_s, \tag{4.12}$$

$$x_{isr}^t,\ y_{isr}^t,\ w_{sr}^t \in \{0,1\},\ \forall i\in\mathcal{I},\ \forall s\in\mathcal{S},\ \forall r\in\mathcal{R}_s,\ \forall t\in\{0,\cdots,K-1\}, \tag{4.13}$$

$$z_s^t,\ f_{is}\ g_{sr}\ h_{sr} \geq 0,\ \forall s\in\mathcal{S},\ \forall r\in\mathcal{R}_s,\ \forall t\in\{0,\cdots,K-1\} \tag{4.14}$$

The objective function (4.1) is to minimize the overtime cost in both OR and PACU and the amortized capacity construction cost of them. Note that R_s, $\forall s \in \mathcal{S}$, is a parameter. Constraint (4.2) ensures that for each stage s, each patient i starts his stay in only one room r. Constraint

(4.3) indicates that for each stage s, at most one patient can be processed at any time in any room. Constraint (4.4) provides p_{is} consecutive time slots for patient i in stage s. Constraint (4.5) counts the number of occupied rooms at time slot t in stage s. Constraint (4.6) assures that the total duration times equal to the total number of time slots that the rooms are occupied in stage s over the entire planning horizon. Constraint (4.7) defines completion time of patient i in stage s. Specifically, the completion time equals to the sum starting time of patient i and her processing time. Constraint (4.8) ensures that patient i in OR cannot proceed to PACU until finishing execution is OR. Constraint (4.9) ensures that there is no-wait between OR and PACU. Constraints (4.10) and (4.11) collectively ensure that the room completion time is nonzero as long as any patient occupies the room. Constraint (4.12) defines the overtime amount. Constraints (4.13) and (4.14) denote the binary and non-negative constraints, respectively.

Administrators of healthcare providers prefer a block-scheduling approach that develops schedules for a group of patients rather than for an individual patient or a service type because of its flexibility in generating schedules (Price et al., 2011). In this context, we also define the duration time of patient i in stage s, p_{is}, based on the patient group in which the patient i is assigned (as discussed in Section 4.4). As provided in Appendix C.3, p_{is} in our data follows log-normal or weibull distribution, as in the literature (e.g., Gul et al., 2011). Later, in the computational study, we extend our model to study to what extent the duration uncertainty affects the capacity.

In Lemma 1 below, we first formally show that Algorithm *FFP1* derives the upper bound of the number of ORs and PACUs for Problem *P1*.

Lemma 1. *Algorithm FFP1 provides an upper bound for the objective of Problem P1 and on the number ORs and PACUs.*

Proof: Algorithm *FFP1* provides a solution to Problem *P1* with zero overtime cost as it obtains a schedule with all completion times before time T, except for the overnight-stay patients in PACU. The solution is feasible to *P1* and thus it is an upper bound for the objective of Problem *P1* and for the numbers of OR and PACU. The total cost of this solution is π, where $\pi = \Lambda C_1^e + (\Lambda + n_o) C_2^e$

+ {the cost of overnight-stay at PACU}. It is easy to see that *FFP1* obtains solution with an upper bound on the number of ORs and PACUs. The reasoning is as follows: if the number of ORs (respectively, PACUs) is set to $\Lambda + 1$ (respectively, $\Lambda + n_o + 1$), then the objective function value increases above π as the cost of deploying more rooms increases and the overtime cost remains zero. Also, the cost of overnight-stay at PACU remains the same in the optimal solution. Thus, *FFP1* provides upper bounds on the number of ORs and PACUs. □

Minimizing the makespan in the two-stage hybrid flow shop problem with no-wait between the two processing stages is known to be strongly NP-hard, even in the case in which stage one contains one machine, $m_1 = 1$, and stage 2 has two machines, $m_2 = 2$ (denoted as $F_2(m_1, m_2)|$no-wait, $m_1 = 1,\ m_2 = 2|C_{max}$) (Sriskandarajah and Ladet, 1986). This complexity result can easily be carried over to the decision version of Problem *P1*. Hence, it is strongly NP-hard, even when $R_2 = 2$, $\forall s \in \mathcal{S}$, as shown in Theorem 1.

Theorem 1. *The decision problem corresponding to Problem P1 is strongly NP-complete, even when $R_1 \geq 2$ and $R_2 \geq 2$.*

Proof: Consider the decision problem devised in Sriskandarajah and Ladet (1986) for problem $F_2(m_1, m_2)|$ no-wait, $m_1 = 1,\ m_2 = 2|C_{max}$, where the number of jobs, $n = 5N + 2$, $m_1 = 1$, $m_2 = 2$, and the makespan, $D \leq FT$, where $FT = (12N + 1)L + (N + 2)B$, where the integers N and B are the parameters from 3-Partition and $L = (N + 3)B$. Now we construct the decision problem for Problem *P1* with $R_1 = 1$ and $R_2 = 2$ as follows: Set the length of regular hours, $T = FT = (12N + 1)L + (N + 2)B$ and the number of patients, $I = n = 5N + 2$. We now need to answer the following question: Does there exist a sequence of patients such that the total cost π is less than or equal to $C_1^e + 2C_2^e$? Note that π does not include overtime cost. It is easy to see that there exists a solution to Problem *P1* with $\pi \leq C_1^e + 2C_2^e$ only if there exists a solution to problem $F_2(m_1, m_2)|$no-wait, $m_1 = 1,\ m_2 = 2|C_{max}$ with the makespan, $D \leq (12N + 1)L + (N + 2)B$, and vice-versa. This completes the proof. □

As discussed earlier, a no-wait constraint is a special case of a blocking constraint. We adopt the no-wait constraint in our main model because of its practical relevance in ASC operations. However, we show in Theorem 2 that Problem *P1* with blocking constraint is indeed equivalent to that with no-wait constraint.

Theorem 2. *Problem P1 with blocking constraint is equivalent to Problem P1 with no-wait constraint for* $R_1 \geq 1$ *and* $R_2 \geq 1$.

Proof: Each feasible schedule to Problem *P1* with blocking constraint can easily be converted into a schedule for Problem *P1* with no-wait constraint with the same completion times for all ORs and PACUs, by changing the processing times and vice-versa. Thus, the result follows because there is no change in total cost as a result of this transformed schedule. □

Along with overtime, idle time in ORs and PACUs is often discussed in patient scheduling literature (e.g., Bai, 2017). In Theorem 3, we show that Problem *P1* is equivalent to the problem that also including idle time cost in ORs and PACUs if certain conditions hold. The condition $\frac{C_1^o}{C_1^d} = \frac{C_2^o}{C_2^d}$ in Theorem 3 is a generalized expression for, for example, $C_1^o = 1.7C_1^d$ and $C_2^o = 1.7C_2^d$, which are often claimed in the literature (Jebali et al., 2006; Lamiri et al., 2008; Pulido et al., 2014), i.e., the overtime cost is around 70% higher than the idle time cost for both stages.

Theorem 3. *Problem P1 is equivalent to Problem* $\hat{P}_1$ *(*$\hat{P}_1$ *includes cost of idle time incurred in OR and PACU beds) when* $C_1^o > C_2^o > 0$*,* $C_1^o > C_1^d > C_2^d > 0$*, and* $\frac{C_1^o}{C_1^d} = \frac{C_2^o}{C_2^d}$*, where* C_s^d *denotes unit cost of bed idle time in stage* s*,* $s = 1, 2$.

Proof: Note that the numbers of ORs and PACUs are fixed at R_1 and R_2, respectively. Let σ (respectively, $\hat{\sigma}$) be an optimal schedule for Problem *P1* (respectively, Problem $\hat{P}_1$). T is the regular operating time. Overtime occurs if OR and PACU are used beyond time T. Thus the regular operating time available for OR (respectively, PACU) is R_1T (respectively, R_2T).
Let t_s^o (respectively, $\hat{t_s^o}$) be the total overtime in σ (respectively, $\hat{\sigma}$), where $s = 1, 2$.
Let t_s^d (respectively, $\hat{t_s^d}$) be the total idle time in σ (respectively, $\hat{\sigma}$), where $s = 1, 2$.
Note that $R_sT - t_s^d + t_s^o = R_sT - \hat{t_s^d} + \hat{t_s^o}$, where $s = 1, 2$. This implies that $\hat{t_s^d} = t_s^d + \hat{t_s^o} - t_s^o$.

Let π_σ (respectively, $\pi_{\hat{\sigma}}$) account for the total idle and overtime cost in σ (respectively, $\hat{\sigma}$). Note the bed costs are the same in σ and $\hat{\sigma}$. Thus, we have

$\pi_\sigma = \sum_s t_s^d C_s^d + \sum_s t_s^o C_s^o$ and $\pi_{\hat{\sigma}} = \sum_s \hat{t_s^d} C_s^d + \sum_s \hat{t_s^o} C_s^o$.

Note that $\pi_{\hat{\sigma}} = \sum_s \hat{t_s^d} C_s^d + \sum_s \hat{t_s^o} C_s^o = \sum_s (t_s^d + \hat{t_s^o} - t_s^o) C_s^d + \sum_s \hat{t_s^o} C_s^o$, That is,

$\pi_{\hat{\sigma}} = (\sum_s t_s^d C_s^d + \sum_s t_s^o C_s^o) + \sum_s (\hat{t_s^o} - t_s^o) C_s^d + \sum_s (\hat{t_s^o} - t_s^o) C_s^o$. This implies that

$\pi_{\hat{\sigma}} = \pi_\sigma + \sum_s (\hat{t_s^o} - t_s^o) C_s^d + \sum_s (\hat{t_s^o} - t_s^o) C_s^o = \pi_\sigma + \sum_s (\hat{t_s^o} - t_s^o)(C_s^d + C_s^o)$.

Note that $\hat{t_1^o} C_1^o + \hat{t_2^o} C_2^o \geq t_1^o C_1^o + t_2^o C_2^o$, otherwise it contradicts that σ is optimal for *P1*. Moreover, note that $\sum_s (\hat{t_s^o} - t_s^o)(C_s^d + C_s^o) \leq 0$, otherwise it contradicts that $\hat{\sigma}$ is optimal for $\hat{\mathbf{P}}_{\mathbf{1}}$. This implies that $(\hat{t_1^o} - t_1^o)(C_1^d + C_1^o) + (\hat{t_2^o} - t_2^o)(C_2^d + C_2^o) \leq 0$. Thus, we have

$$\hat{t_1^o} C_1^o + \hat{t_2^o} C_2^o \geq t_1^o C_1^o + t_2^o C_2^o. \tag{4.15}$$

$$(\hat{t_1^o} - t_1^o)(C_1^d + C_1^o) + (\hat{t_2^o} - t_2^o)(C_2^d + C_2^o) \leq 0. \tag{4.16}$$

The results follows from the four cases below:

Case 1: $\hat{t_1^o} \leq t_1^o$ and $\hat{t_2^o} \leq t_2^o$. This implies that $\hat{t_1^o} C_1^o + \hat{t_2^o} C_2^o \leq t_1^o C_1^o + t_2^o C_2^o$. Together with relation (4.16) above, we have $\hat{t_1^o} C_1^o + \hat{t_2^o} C_2^o = t_1^o C_1^o + t_2^o C_2^o$. This in turn implies that $\hat{t_1^o} = t_1^o$ and $\hat{t_2^o} = t_2^o$, and thus $\sigma = \hat{\sigma}$.

Case 2: $\hat{t_1^o} \leq t_1^o$ and $\hat{t_2^o} > t_2^o$. Let $\hat{t_1^o} + \delta_1 = t_1^o$ and $\hat{t_2^o} = t_2^o + \delta_2$, where $\delta_1 \geq 0$ and $\delta_2 > 0$. Consider two sub-cases:

Case 2.1: $\delta_1 = 0$. Relation (4.16) implies that $\delta_2(C_2^d + C_2^o) \leq 0$. This contradicts with $C_2^d > 0$, $C_2^o > 0$, and $\delta_2 > 0$. Thus, this case is not feasible.

Case 2.2: $\delta_1 > 0$. If $\delta_1 = \delta_2$, then relation (4.16) implies $(\hat{t_2^o} - t_2^o) C_2^o \geq (t_1^o - \hat{t_1^o}) C_1^o \Leftrightarrow \delta_2 C_2^o \geq \delta_1 C_1^o \Leftrightarrow C_2^o \geq C_1^o$, which contradicts with $C_2^o < C_1^o$. If $\delta_1 \neq \delta_2$, Relation (4.16) implies that $\delta_2 C_2^o \geq \delta_1 C_1^o$, i.e., $\frac{\delta_1}{\delta_2} \leq \frac{C_2^o}{C_1^o}$. Relation (4.16) implies that $\delta_2(C_2^d + C_2^o) \leq \delta_1(C_1^d + C_1^o)$, that is $\frac{\delta_1}{\delta_2} \geq \frac{C_2^d + C_2^o}{C_1^d + C_1^o}$. This case is feasible if $\frac{C_2^o}{C_1^o} \geq \frac{C_2^d + C_2^o}{C_1^d + C_1^o} \Leftrightarrow C_2^o(C_1^d + C_1^o) \geq C_1^o(C_2^d + C_2^o)$, that is $C_2^o C_1^d \geq C_1^o C_2^d$, or equivalently, $\frac{C_1^o}{C_1^d} \leq \frac{C_2^o}{C_2^d}$.

Case 3: $\hat{t_1^o} > t_1^o$ and $\hat{t_2^o} \leq t_2^o$. Similar to Case 2, let $\hat{t_1^o} = t_1^o + \delta_1$ and $\hat{t_2^o} + \delta_2 = t_2^o$, where $\delta_1 > 0$ and $\delta_2 \geq 0$. Consider two sub-cases:

Case 3.1: $\delta_2 = 0$. Relation (4.16) implies that $\delta_1(C_1^d + C_1^o) \leq 0$. This contradicts with $C_1^d > 0$, $C_1^o > 0$, and $\delta_2 > 0$. Thus, this case is not feasible.

Case 3.2: $\delta_2 > 0$. If $\delta_2 = \delta_1$, then Relation (4.16) implies $(\hat{t_1^o} - t_1^o)(C_1^d + C_1^o) \leq (t_2^o - \hat{t_2^o})(C_2^d + C_2^o) \Leftrightarrow \delta_1(C_1^d + C_1^o) \leq \delta_2(C_2^d + C_2^o) \Leftrightarrow C_1^d + C_1^o \leq C_2^d + C_2^o$, which contradicts that $C_1^d > C_2^d$ and $C_1^o > C_2^o$. If $\delta_2 \neq \delta_1$, Relation (4.16) implies that $\delta_1(C_1^d + C_1^o) \leq \delta_2(C_2^d + C_2^o)$, i.e., $\frac{\delta_1}{\delta_2} \leq \frac{C_2^d + C_2^o}{C_1^d + C_1^o}$. Relation (4.16) implies $(t_2^o - \hat{t_2^o})C_2^o \leq (\hat{t_1^o} - t_1^o)C_1^o \Leftrightarrow \delta_2 C_2^o \leq \delta_1 C_1^o$, i.e., $\frac{\delta_1}{\delta_2} \geq \frac{C_2^o}{C_1^o}$. This case is feasible if $\frac{C_2^d + C_2^o}{C_1^d + C_1^o} \geq \frac{C_2^o}{C_1^o} \Leftrightarrow C_1^o(C_2^d + C_2^o) \geq C_2^o(C_1^d + C_1^o)$, that is $C_1^o C_2^d \geq C_2^o C_1^d$, or equivalently, $\frac{C_1^o}{C_1^d} \geq \frac{C_2^o}{C_2^d}$. From Case 2.2 and Case 3.2, $\sigma = \hat{\sigma}$ when $\frac{C_1^o}{C_1^d} = \frac{C_2^o}{C_2^d}$.

Case 4: $\hat{t_1^o} > t_1^o$ and $\hat{t_2^o} > t_2^o$. This implies that $\hat{t_1^o} C_1^o + \hat{t_2^o} C_2^o > t_1^o C_1^o + t_2^o C_2^o$. By rearranging the terms, we obtain $(\hat{t_1^o} - t_1^o)(C_1^d + C_1^o) + (\hat{t_2^o} - t_2^o)(C_2^d + C_2^o) > 0$ that contradicts with Relation (4.16). Hence, this case is not feasible. □

In view of the NP-completeness in the strong sense of Problem *P1*, developing an efficient and effective heuristic is desirable. We propose a heuristic, namely *BackwardASC* in Section 4.5.2.4, that first generates a daily patient schedule in OR and PACU stages. In the next section, we propose an algorithm that derives the optimal numbers of OR and PACU.

4.5.2.2 Algorithm AdaptiveASC to Derive Optimal Numbers of OR and PACU

For a given set of patients, I^w, in each weekday $w \in W = \{1, \cdots, 5\}$, Problem *P1* minimizes total cost, namely the sum of overtime cost and construction cost, when the number of rooms R_s in each stage s is a fixed parameter. Motivated from the trade-off between the two cost measures, we propose an algorithm, namely *AdaptiveASC*, to find the optimal numbers of ORs and PACUs that minimize the total cost to serve patient demand over weekdays. The main idea is to efficiently update the number of rooms in each stage after initializing the number of ORs as $R_1 = \max_{w \in W}\{\Lambda^w\}$ and PACUs as $R_2 = \max_{w \in W}\{\Lambda^w + n_o^w\}$, where Λ^w is the number of ORs derived by Algorithm *FFP1* and n_o^w is the number of overnight-stay patients in weekday w.

Let $\Pi(r_1, r_2)$ be the sum of objective function values of Problem *P1* over weekdays, $\forall w \in W$, where r_1 and r_2 represent the numbers of ORs and PACUs, respectively. Let $\Pi(r_1, r_2) = \Pi^o(r_1, r_2) + \Pi^e(r_1, r_2)$, where Π^o and Π^e denote the overtime cost and the construction cost, respectively. We include a visual representation of Algorithm *AdaptiveASC* in Figure 4.4. $\Pi(r_1, r_2)$ is convex over $r_1 \in [1, R_1]$ and $r_2 \in [1, R_2]$ as illustrated later in Figure 4.6 in the computational study . Lemma 2 shows that there indeed exists r_2^* such that $\Pi(r_1, r_2^*)$ is less than equal to $\Pi(r_1, r_2^* - 1)$ and $\Pi(r_1, r_2^* + 1)$, given r_1 is fixed. Similar case can be shown for r_1 when r_2 is fixed.

Lemma 2. *Given the number of ORs, R_1, there exists r_2 such that the following relations hold:*

1. $\Pi(R_1, r_2 + 1) \geq \Pi(R_1, r_2)$*, and*

2. $\Pi(R_1, r_2 - 1) \geq \Pi(R_1, r_2)$*,*

where $\Pi(r_1, r_2)$ is an optimal solution obtained by solving Problem P1 for given r_1 and r_2.

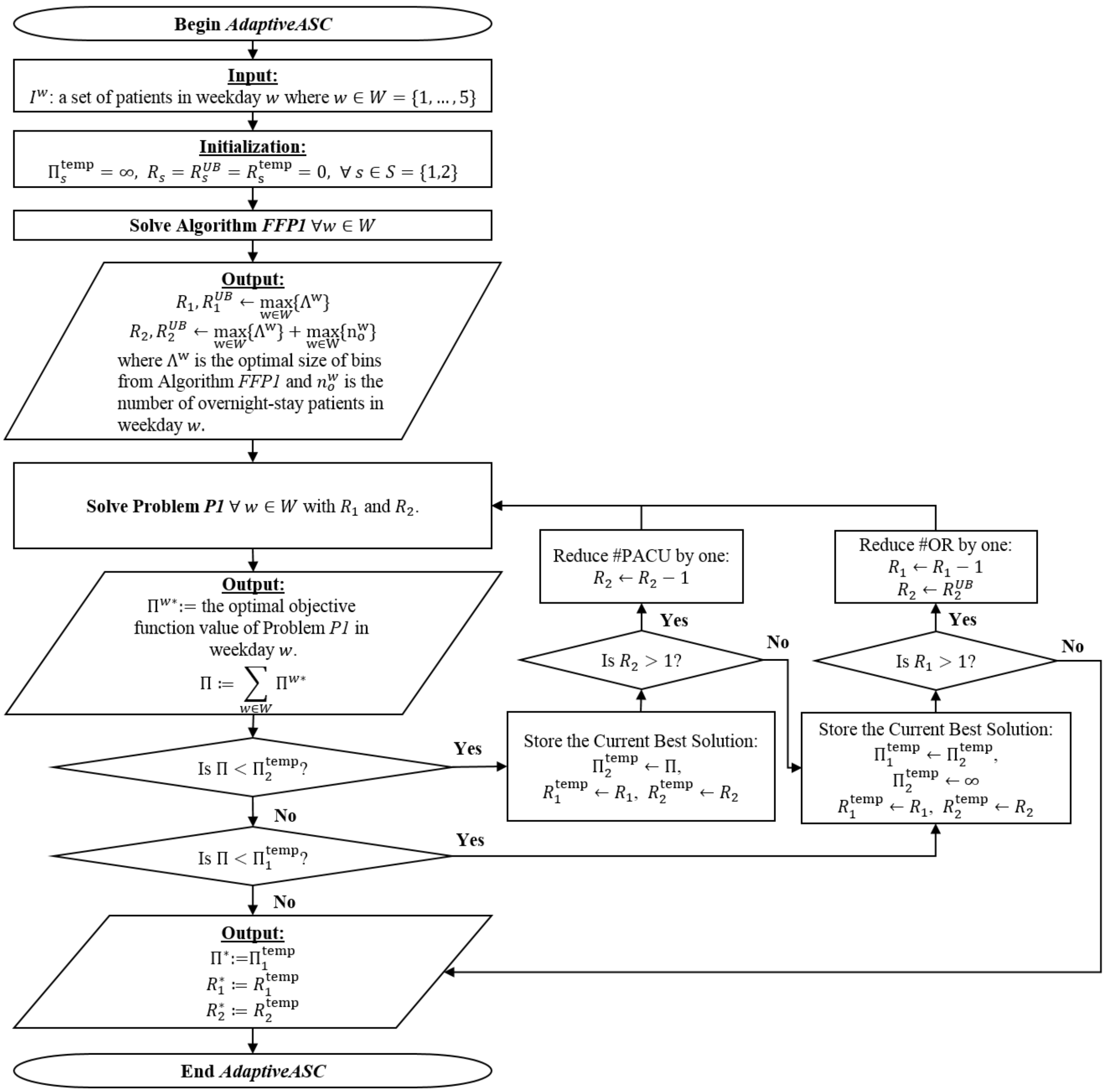
Proof: The results follow from the trade-off between the amortized PACU construction cost C_2^e and the PACU overtime cost C_2^o. As long as $\Pi^o(R_1, k) - \Pi^o(R_1, k+1) < C_2^e$, it is beneficial to reduce the number of PACUs (Item 1), because Π^o is a non-increasing function of r_2. Hence, for a certain number of PACU r_2, $\Pi^o(R_1, r_2 - 1) - \Pi^o(R_1, r_2) \geq C_2^e$ holds (Item 2). □

Algorithm *AdaptiveASC* terminates after a finite number of iterations as shown in Lemma 3.

Lemma 3. *Algorithm AdaptiveASC terminates in a finite number of iterations and its time complexity is $O(\Lambda^2)$.*

Proof: We begin the algorithm with the initial numbers of ORs and PACUs, namely $R_1^{UB} = \max_{w \in W}\{\Lambda^w\}$ and $R_2^{UB} = \max_{w \in W}\{\Lambda^w\} + \max_{w \in W}\{n_o^w\}$, obtained by Algorithm *FFP1* and the number of overnight-stay patients. Algorithm *AdaptiveASC* updates either the number of ORs or PACUs and evaluates the objective function of Problem *P1*. Specifically, beginning with the initial number of ORs, R_1^{UB}, we calculate the total cost starting from the initial number of PACUs, R_2^{UB}. Next, we obtain the total cost after reducing the number of PACUs by one. We continue this process

Figure 4.4: Flow Chart of Algorithm AdaptiveASC



as long as the total cost decreases. Once the total cost begins to increase, we store the currently minimum total cost and stop reducing the number of PACU. Instead, now we reduce the number of OR by one, reset the number of PACU to R_2^{UB}, and again calculate the total cost by reducing the number of PACU by one in each iteration. In this way, we search the two-dimensional space until there is no improvement in Π and obtain the optimal numbers of ORs and PACUs. Since the numbers of OR and PACU can be integers in $[1, R_1^{UB}]$ and $[1, R_2^{UB}]$, respectively, the algorithm terminates in a finite number of iterations and we get quadratic time complexity, i.e., $O(N^2)$ where $N = \max_{w \in W}\{\Lambda^w\}$. □

Further, Theorem 4 shows that Algorithm *AdaptiveASC* find an optimal solution in the finite number of iterations.

Theorem 4. *Algorithm AdaptiveASC terminates with an optimal solution* $\Pi(r_1^*, r_2^*)$ *that minimize the sum of capacity construction cost and overtime cost, where* $\Pi(r_1, r_2)$ *is obtained by solving Problem P1 optimally for given* r_1 *and* r_2.

Proof: It is sufficient to prove that Algorithm *AdaptiveASC* finds an optimal solution for given R_1. Suppose the optimal solution is $\Pi(R_1, \bar{r}_2)$ and Algorithm *AdaptiveASC* terminates with a solution $\Pi(R_1, \hat{r}_2)$ such that $\Pi(R_1, \bar{r}_2) < \Pi(R_1, \hat{r}_2)$. Then, $\Pi(R_1, \bar{r}_2) > \Pi(R_1, \bar{r}_2 - 1)$ or $\Pi(R_1, \bar{r}_2) > \Pi(R_1, \bar{r}_2 + 1)$ must hold, otherwise the solution $\Pi(R_1, \bar{r}_2)$ would have been found by the algorithm as shown in Lemma 2. These relations contradict the optimality of $\Pi(R_1, \bar{r}_2)$. Hence, the result follows. □

In Algorithm *AdaptiveASC*, iteratively solving Problem *P1* is computationally challenging. We thus develop an efficient and effective heuristic, namely *BackwardASC*. Before, we derive lower bounds on Problem *P1* that we use to evaluate the heuristic.

4.5.2.3 Lower Bound on the Makespan and the Total Cost

The majority of scheduling literature discusses lower bounds on the makespan but not on the cost objective function (Ruiz and Vázquez-Rodríguez, 2010). We also recognize a significant challenge in specifying a proper lower bound on our cost objective function. However, since we are only able to solve Problem *P1* for small-sized instances, e.g., up to ten patients in a day, we propose simple lower bounds on both the makespan and the objective function for Problem *P1*. We use these lower bounds to evaluate the performance of Heuristic *BackwardASC* for large-sized problem instances.

Consider the problem of scheduling a set of independent patients $\mathcal{I} = \{1, 2, \cdots, I\}$, in which each patient i consists of a chain of two operations .$\{O_{i1}, O_{i2}\}$, $\forall i \in \mathcal{I}$ The first operation O_{i1} of any patient i needs to be processed on the rooms in stage 1 during an uninterrupted processing time $p_{i1} \geq 0$, and then the second operation O_{i2} needs to be processed on one of the rooms in stage 2 during an uninterrupted processing time $p_{i2} \geq 0$. We denote R_1 parallel ORs in the first stage by $\{M_{1,1}, M_{1,2}, \cdots, M_{1,R_1}\}$ and the R_2 parallel PACUs in the second stage by $\{M_{2,1}, M_{2,2}, \cdots, M_{2,R_2}\}$.

Let f_m^H (respectively, f_c^H) denote the makespan (respectively, the total cost) obtained by any heuristic H and use f_m^{LB} (resp., f_c^{LB}) to stand for a lower bound of the optimal makespan (respectively, the optimal total cost). We use the following formulas to calculate the lower bounds:

$$f_m^{LB} = \max\Big\{\frac{1}{R_1}\sum_{i\in\mathcal{I}} p_{i1} + \min_{i\in\mathcal{I}} p_{i2}, \frac{1}{R_2}\sum_{i\in\mathcal{I}} p_{i2} + \min_{i\in\mathcal{I}} p_{i1}, \max_{i\in\mathcal{I}}\{p_{i1} + p_{i2}\}\Big\}. \quad (4.17)$$

$$\begin{aligned} f_c^{LB1} =& \min_{s\in\mathcal{S}} C_s^o \cdot \max\Big\{0, R_1 \cdot \Big(\frac{1}{R_1}\sum_{i\in\mathcal{I}} p_{i1} + \min_{i\in\mathcal{I}} p_{i2} - T\Big), \\ & R_2 \cdot \Big(\frac{1}{R_2}\sum_{i\in\mathcal{I}} p_{i2} + \min_{i\in\mathcal{I}} p_{i1} - T\Big)\Big\} + \sum_{s\in\mathcal{S}} C_s^e R_s \end{aligned} \quad (4.18)$$

Additionally, we propose two alternative lower bounds for the total cost. The idea on the second lower bound f_c^{LB2} begins from the optimal total cost f_c^* that can be stated as follows:

$$f_c^* = C_1^o \cdot \max\Big\{0, \sum_{i\in\mathcal{I}} p_{i1} + t_1^d - R_1 T\Big\} + C_2^o \cdot \max\Big\{0, \sum_{i\in\mathcal{I}} p_{i2} + t_2^d - R_2 T\Big\} + \sum_{s\in\mathcal{S}} C_s^e R_s \quad (4.19)$$

where t_1^d and t_2^d are the idle times in the first stage and the second stage, respectively, in the optimal solution (as defined in Theorem 3). Since $t_1^d \geq 0$ and $t_2^d \geq R_2 \cdot \min_{i\in\mathcal{I}} p_{i1}$, we have:

$$f_c^{LB2} = C_1^o \cdot \max\left\{0, \sum_{i\in\mathcal{I}} p_{i1} - R_1 T\right\} + C_2^o \cdot \max\left\{0, \sum_{i\in\mathcal{I}} p_{i2} + R_2 \cdot \min_{i\in\mathcal{I}} p_{i1} - R_2 T\right\} + \sum_{s\in\mathcal{S}} C_s^e R_s \leq f_c^*$$

The third lower bound f_c^{LB3} better captures the overtime caused by overnight-stay patients. For any patient i, her earliest time to enter any PACU is p_{i1}. If the patient stays overnight, then we can capture the PACU overtime of her as $p_{i1} + p_{i2} - T$. Hence, we have:

$$f_c^{LB3} = C_1^o \cdot \max\left\{0, \sum_{i\in\mathcal{I}} p_{i1} - R_1 T\right\} + C_2^o \cdot \sum_{i\in\mathcal{I}} \max\left\{0, p_{i1} + p_{i2} - T\right\} + \sum_{s\in\mathcal{S}} C_s^e R_s \tag{4.20}$$

The lower bound on total cost used to evaluate heuristics is $f_c^{LB} = \max\{f_c^{LB1}, f_c^{LB2}, f_c^{LB3}\}$ (Section 4.6.1). We use the relative percentage above the lower bound, Gap(%) $= \frac{100(f_\cdot^H - f_\cdot^{LB})}{f_\cdot^{LB}}$ as the performance measure of heuristic H for both f_m^{LB} and f_c^{LB}.

4.5.2.4 Heuristic BackwardASC to Solve Problem P1

Heuristic *BackwardASC* aims to assign patients to an ASC that has R_1 ORs and R_2 PACUs to minimize the overtime cost. Notice that R_1 is not necessarily equal to R_2. The main idea of heuristic *BackwardASC* is to tightly sequence patients into ORs and PACUs to reduce idle times, and thus to minimize the overtime in the system, while obeying the no-wait constraint between the two stages. Building upon a minimum deviation algorithm suggested by Xie et al. (2004) that aims to minimize makespan, heuristic *BackwardASC* splits the patients into two groups, i.e., (1) overnight-stay and (2) non-overnight-stay. The heuristic first sequences the overnight-stay patients to prevent excessive overtime in PACU caused by the overnight-stay patients. Once all overnight-stay patients are scheduled, we schedule non-overnight-stay patients in the same approach. Heuristic *BackwardASC* is a greedy algorithm, i.e., once a partial schedule is obtained, this partial schedule will not be changed.

Each loop schedules one patient. At the beginning of each loop, let M_{1,k_1} be the OR that first becomes idle, and let t_1 be the time when M_{1,k_1} becomes idle. Similarly, let M_{2,k_2} be the PACU that is the first idle room among other PACUs, and let t_2 be the time that M_{2,k_2} becomes idle. We find an unscheduled patient i with p_{i1} being closest to $\max\{0, t_2 - t_1\}$ (i.e., $i = \text{argmin}_{i\in\mathcal{I}}|p_{i1} - \max\{0, t_2 - t_1\}|$). To reduce the idle time, patient i is scheduled next on the first stage OR M_{1,k_1} and the second stage PACU M_{2,k_2}. In other words, if $p_{i1} \geq t_2 - t_1$, patient i starts to process on OR M_{1,k_1} immediately; otherwise, patient i has to start on OR M_{1,k_1} after $(t_2 - t_1 - p_{1i})$ amount of waiting time. By doing so, we iteratively generate a new partial schedule with one more patient scheduled.

Heuristic *BackwardASC* for Problem *P1* (*BackwardASC*)

Input: There are R_1 ORs, $\{M_{1,1}, M_{1,2}, ..., M_{1,R_1}\}$, and R_2 PACUs, $\{M_{2,1}, M_{2,2}, ..., M_{2,R_2}\}$. $\mathcal{I} = \{1, 2, ..., I\}$ is a set of patients to be scheduled. $\mathcal{I} = \mathcal{I}^o \cup \mathcal{I}^{no}$, where $\mathcal{I}^o$ is the set of overnight-stay patients and $\mathcal{I}^{no}$ is the set of non-overnight-stay patients. (p_{1i}, p_{2i}) is the duration times of patient $i \in \mathcal{I}$ in (OR, PACU).

Step 1: Let M_{1,k_1} be the OR that first becomes idle (break ties arbitrarily), and let t_1 be the time when M_{1,k_1} becomes idle. Similarly, let M_{2,k_2} be the PACU that first becomes idle (break ties arbitrarily), and let t_2 be the time when M_{2,k_2} becomes idle.

Step 2: If $\mathcal{I}^o \neq \emptyset$, find an unscheduled patient $i \in \mathcal{I}^o$ with p_{i1} being closest to $\max\{0, t_2 - t_1\}$; otherwise, find an unscheduled patient $i \in \mathcal{I}^{no}$ with p_{i1} being closest to $\max\{0, t_2 - t_1\}$.

Step 3: To reduce the idle time, schedule patient i next on the first stage OR M_{1,k_1} and the second stage PACU M_{2,k_2}. In other words, if $p_{i1} \geq t_2 - t_1$, patient i starts to process on OR M_{1,k_1} immediately; otherwise, patient i starts on OR M_{1,k_1} after waiting $(t_2 - t_1 - p_{1i})$ time.

Step 4: If $\mathcal{I} = \emptyset$, stop. Otherwise, go to Step 1.

Output: A schedule for patients in $\mathcal{I}$.

We provide a formal description of Heuristic *BackwardASC* in Appendix C.4 and its numerical illustration in Appendix C.5. We later evaluate the performance of *BackwardASC* in the computational study. Specifically, we compare the effectiveness of the heuristic with Problem *P1* using small instances and the lower bounds discussed in Section 4.5.2.3.

4.5.3 Obtain the Minimal Number of HRs

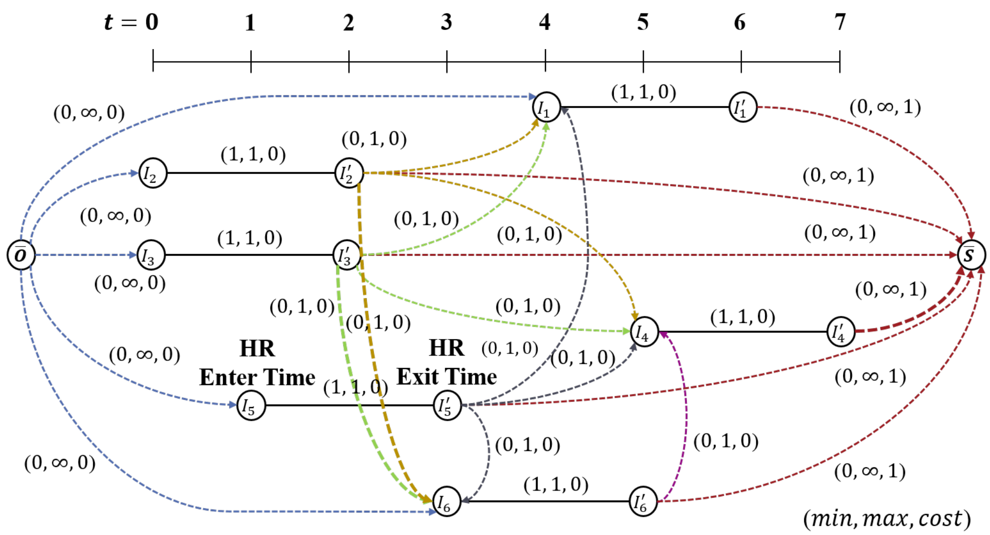
Having obtained the optimal numbers of OR and PACU and the patient schedules, we now derive the minimal number of HRs that preserve the previously obtained schedule of patients in ORs and PACUs. Here, we consider HRs as buffers that cost much less than ORs and PACUs, thus cost measures related to HRs are ignored.

The problem can be formulated as a minimum cost flow (MCF) problem. In our construction of the MCF network, nodes I_i and $I_i^{'}$ correspond to HR enter time and HR exit time of patient i. $\bar{O}$ and $\bar{S}$ denote a source node and a sink node, respectively. The upper and lower bounds on an arc $(I_i, I_i^{'})$ are 1, enforcing each patient to be served by exactly one HR. The arc $(I_i^{'}, \bar{S})$ linking the patient HR exit time and sink node has a unit flow cost. Figure 4.5 illustrates a simple example of the MCF model with six patients. In this example, we need at least three HRs to preserve the HR schedule of the six patients (e.g., patients 1 and 2 in HR1, patients 3 and 4 in HR2, patients 5 and 6 in HR3), and thus the solution is three. The MCF admits a polynomial-time solution as shown in Lemma 4.

Lemma 4. *The MCF model admits a polynomial-time solution and is equivalent to optimizing the minimum number of HRs that preserve the sequence of patients in ORs. The overall complexity of MCF is* $O\left(n^4 \log n\right)$*, where* $n = 2(I + 1)$.

Proof: The goal of the MCF model is to serve patients with minimum number of HRs while preserving the sequence of the patients in ORs. It is straightforward to verify that each feasible flow corresponds to a feasible pre-operative service plan in a single HR. Thus, the optimal solution to the MCF model constructed above is equivalent to optimally deciding the number of HRs, and the proposed MCF model can be solved polynomially. The complexity of solving the MCF model is $O(n^4 \log n)$, where $n = 2(I+1)$ is the number of nodes in the network (Ahuja et al., 1993). The result holds in general. □

Figure 4.5: Minimum Cost Flow diagram for determining the minimum number of HRs that preserve the sequence of patients in ORs [$\bar{O}$: Source, I: Patients, $\bar{S}$: $Sink$]



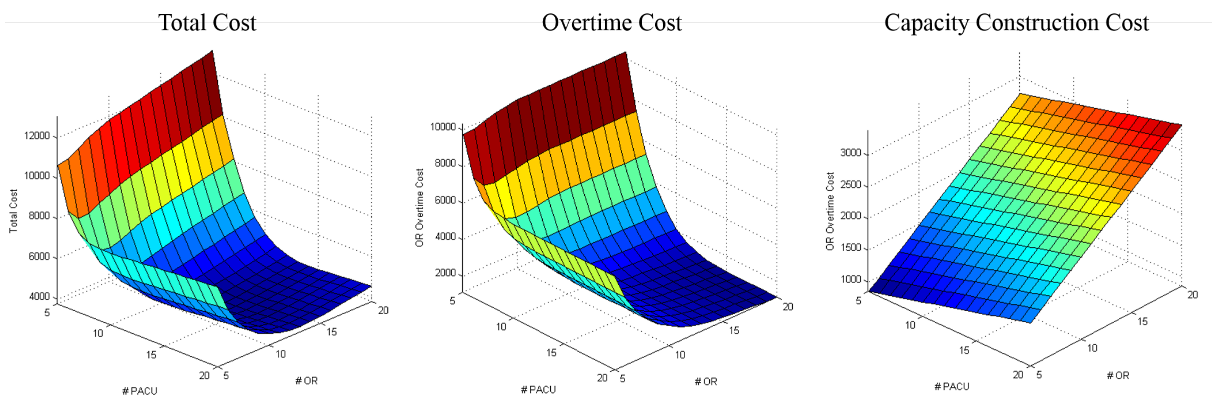
Notes: Solid lines indicate flows through HRs.

4.6 Computational Experiments

The key objectives of our computational study are (i) to evaluate the performance of Heuristic *BackwardASC* with respect to Problem P1 (Section 4.6.1), and (ii) to explore the impact of various business parameters related to ASC operations on capacity (Section 4.6.2). To be more specific, we study the impacts of uncertain composition of patient groups (Section 4.6.2.3), of uncertain patient durations (Section 4.6.2.4), and of uncertain overnight-stays (Section 4.6.2.5). We also conduct sensitivity analysis on the construction and overtime cost parameters (Section 4.6.2.6). For the first objective, we generate a small problem based on the real instances described in Tables 4.5 and 4.6. We also compare the heuristic solution to the lower bound on the objective function. For the second objective, we generate new instances and perform an extensive set of experiments.

Figure 4.6 shows an example of total cost, overtime cost, and amortized capacity construction cost over different combinations of the numbers of ORs and PACUs. In this figure, we observe a convex shape of overtime cost, i.e., increasing overtime cost at an increasing rate as the number of ORs and/or PACUs decreases. The capacity construction cost increases linearly as the numbers of ORs and PACUs increase. Taken together, the total cost function also has a convex shape, similar to EOQ cost curves. This is to be expected, since having excessive capacity leads to a large construction cost even if the overtime cost becomes less. In contrast, having insufficient capacity drastically increases overtime cost with relatively small savings on construction cost. Based on the trade-offs, Algorithm *AdaptiveASC* derives the optimal ORs and PACUs with minimal total cost.

Figure 4.6: Example of Total Cost (left), Overtime Cost (middle), and Amortized Capacity Construction Cost (right) for an Instance in the Computational Study



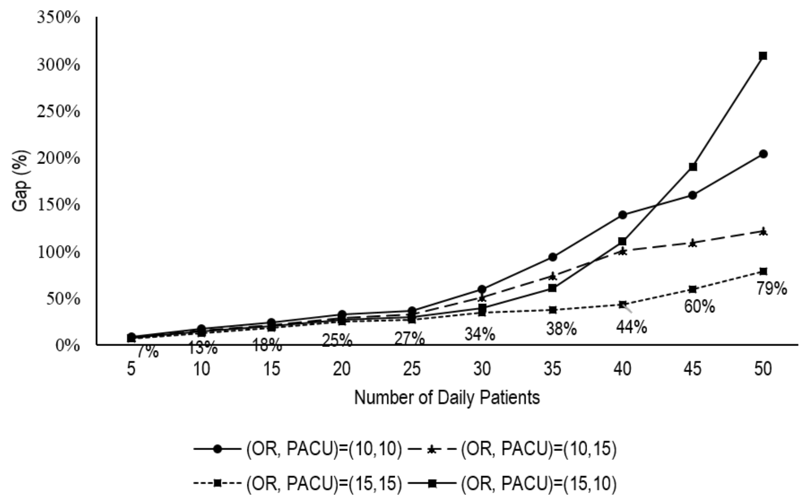
4.6.1 Performance of Heuristic BackwardASC with respect to Problem P1 and its Lower Bound

Due to computational complexity, the optimal solution of Problem *P1* can be obtained only for small-sized instances, e.g., daily patient demand of ten to fifteen. For the instances having optimal solutions[11], the gap between Heuristic *BackwardASC* and Problem *P1* ranges from 4 to 8%. However, such instances are much smaller than the daily number of patients being served in practice. To evaluate the performance of our heuristic for larger-sized instances, we compare total costs relative to the lower bound f_c^{LB} (proposed in Section 4.5.2.3) and show the results in Figure 4.7. If the capacity is too small compared to the daily demand, the gap is large. The gap increases exponentially if the PACU capacity is less then the OR capacity. This observation perhaps indicates

[11]We consider only non-trivial instances that have overtime.

that our lower bounds impose limitations on capturing the extensive overtime in PACUs. However, as long as the number of daily patients is reasonably large for given capacity (i.e., 30 patients for 10 to 15 ORs), the gap is 35% or less. Given limited literature on the cost objective function in the HFS domain (Ruiz and Vázquez-Rodríguez, 2010) and significant challenges to develop well-performing lower bounds, Heuristic *BackwardASC* generates a solution comparable to that of Problem *P1*.

Figure 4.7: Performance of Heuristic BackwardASC with respect to LB



Notes: We randomly generated patient instances with a simulation size of 500 for each capacity combination and the number of daily patients.

4.6.2 Impacts of Business Parameters related to ASC Operations on Capacity

In this section, we first propose different strategies in defining aggregate weekly patient schedules. Afterward, we investigate the impacts of several uncertain business parameters on the capacity decision.

4.6.2.1 Aggregate Weekly Schedule

We provide a guideline for developing a daily patient sample path over weekdays. ASC planners may have a planned weekly patient demand and classify the patients into a small number of patient groups as suggested in Section 4.4. The next step is to allocate the patients over weekdays.

We propose three strategies of defining patient sample paths over weekdays and compare the resulting capacity plans in this computational study. The first strategy is simply to follow the current patient sample paths as listed in Table 4.5. The second strategy is to evenly allocate the patient groups over weekdays, as listed in Table 4.8. Lastly, the third strategy is to define certain weekdays for specific patient groups, as listed in Table 4.9. In these example patient sample paths, we allocate patient groups 1 and 2 (i.e., relatively simple and short surgeries) into Monday and Tuesday and patient groups 3 to 7 (i.e., relatively complex and long surgeries) into the other three weekdays. ASC planners can selectively choose either one of the three strategies or combined version of them, depending on other factors such as the availability of surgeons and equipment.

After performing a numerical study over an extensive set of instances, we observe that the results of the first strategy are quite similar to the second strategy, implying that the current patient sample paths are reasonably balanced over weekdays. Therefore, we focus on comparing the first and third strategies in the rest of the computational studies to provide capacity implications for the two different patient sample path strategies.

Table 4.8: Balanced Patient Sample Path by Weekday (Average # of patients)

Patient Group	Monday	Tuesday	Wedesday	Thursday	Friday
1	11.84 (3.44)	11.84 (3.44)	11.84 (3.44)	11.84 (3.44)	11.84 (3.44)
2	5.52 (1.95)	5.52 (1.95)	5.52 (1.95)	5.52 (1.95)	5.52 (1.95)
3	2.42 (1.24)	2.42 (1.24)	2.42 (1.24)	2.42 (1.24)	2.42 (1.24)
4	8.29 (5.60)	8.29 (5.60)	8.29 (5.60)	8.29 (5.60)	8.29 (5.60)
5	0.55 (0.49)	0.55 (0.49)	0.55 (0.49)	0.55 (0.49)	0.55 (0.49)
6	1.22 (0.41)	1.22 (0.41)	1.22 (0.41)	1.22 (0.41)	1.22 (0.41)
7	2.10 (0.93)	2.10 (0.93)	2.10 (0.93)	2.10 (0.93)	2.10 (0.93)

Notes: Standard deviation of the number of patients is in parenthesis.

Table 4.9: Dedicated Patient Sample Path by Weekday (Average # of patients)

Patient Group	Monday	Tuesday	Wedesday	Thursday	Friday
1	29.60 (8.61)	29.60 (8.61)	0 (0)	0 (0)	0 (0)
2	13.80 (4.88)	13.80 (4.88)	0 (0)	0 (0)	0 (0)
3	0 (0)	0 (0)	4.04 (2.27)	4.04 (2.27)	4.04 (2.27)
4	0 (0)	0 (0)	13.81 (4.48)	13.81 (4.48)	13.81 (4.48)
5	0 (0)	0 (0)	0.91 (1.79)	0.91 (1.79)	0.91 (1.79)
6	0 (0)	0 (0)	2.04 (2.47)	2.04 (2.47)	2.04 (2.47)
7	0 (0)	0 (0)	3.50 (2.32)	3.50 (2.32)	3.50 (2.32)

Notes: Standard deviation of the number of patients is in parenthesis.

4.6.2.2 The Testbed

In our simulation runs, we generate 100 instances for each day. Therefore, in our analysis, we have tested $73,500$ daily instances (i.e., five days a week, three strategies in defining patient sample paths, 16 uncertainty levels on patient group compositions + 16 uncertainty levels on patient durations + 10 probabilities of staying overnight per each patient + 7 different sets of cost parameters: $100 \times 5 \times 3 \times (16 + 16 + 10 + 7)$ instances). The MIP model, Problem *P1*, and related algorithms are implemented in C++ and solved using CPLEX (version 12.6.1) with 2.70 GHz CPU, 32 GB RAM, and Windows 10 operating system. Heuristic *BackwardASC* and related algorithms (e.g., *FFP1* and *AdaptiveASC*) are implemented in Matlab and solved with the same system.

4.6.2.3 Impact of Uncertain Patient Groups Composition on Capacity

We generate a set of patient instances that are characterized by the total number of patients to be served in a day and the composition of patient groups as listed.

Table 4.10 provides a guideline for deciding capacity in each stage for varying levels of uncertainty in patient sample paths over weekdays. When the risk level is γ, the number of patients for each patient group is realized from $(\mu - \gamma\sigma, \mu + \gamma\sigma)$, where μ (respectively, σ) is the mean (respectively, standard deviation) of the number of patients from a focal patient group. For example, if $\gamma = 0.6$, the number of patients classified into Patient Group 2 from dedicated patient sample path strategy (i.e., Table 4.9) is the value of randomly generated number from the range $[13.80 - 0.6 \cdot 4.88, 13.80 + 0.6 \cdot 4.88]$, rounded to the nearest integer.

In Table 4.10, we observe that the capacity in each stage and the total cost increase as the risk level increases, implying that deviating from a daily patient sample path requires additional capacity for both actual and dedicated patient sample path strategies. Interestingly, the dedicated patient sample path strategy costs on average 8.78% more than the actual strategy (with max 17.18% when $\gamma = 2.2$ and min 1.29% when $\gamma = 0.4$) because of excess overtimes driven by the longer surgeries assigned on Wednesday to Friday. Meanwhile, the dedicated strategy requires on average 23% (or 4.1 units) fewer HRs than the actual strategy because the variance of OR durations in each day is smaller. In conclusion, as long as geographical space is not a strict constraint (especially for HRs), the actual patient sample path strategy is recommended as it generally leads to lower total costs.

Table 4.10: Results of Computational Experiment: Uncertain Composition of Patient Groups

	Actual Patient Sample Path						Dedicated Patient Sample Path					
Risk Lv.	#HR	#OR	#PACU	OR OT Cost	PACU OT Cost	Total Cost	#HR	#OR	#PACU	OR OT Cost	PACU OT Cost	Total Cost
0.0	16.0	8	10	118.6	947.8	2516.3	11.2	8	11	109.0	1056.8	2660.8
0.2	16.0	8	10	122.7	930.8	2503.6	11.2	8	11	96.8	956.8	2548.6
0.4	16.0	8	10	153.4	980.1	2583.5	11.2	8	11	125.5	996.9	2617.4
0.6	16.0	8	10	169.7	981.0	2600.8	12.6	9	11	118.5	956.9	2695.4
0.8	18.0	9	11	102.1	926.1	2648.2	12.6	9	11	125.9	950.2	2695.4
1.0	17.9	9	11	123.5	940.0	2683.5	12.6	9	12	173.7	1003.7	2842.4
1.2	17.8	9	11	164.4	1014.7	2799.1	12.6	9	13	186.9	1058.9	2955.8
1.4	17.7	9	11	184.3	1006.7	2811.0	13.9	10	14	152.2	1030.1	3062.3
1.6	17.6	9	12	194.6	1072.1	2931.7	13.8	10	14	218.4	1097.2	3195.6
1.8	17.5	9	12	218.7	1105.8	2989.4	15.0	11	14	214.5	1062.0	3281.4
2.0	19.2	10	12	181.2	1054.2	3025.4	14.9	11	15	242.7	1161.2	3453.9
2.2	18.9	10	13	222.2	1113.9	3171.1	16.1	12	17	282.0	1282.3	3829.3
2.4	18.8	10	13	266.2	1163.2	3264.4	16.0	12	17	318.1	1306.7	3889.9
2.6	20.5	11	14	238.1	1217.1	3460.2	17.3	13	18	308.6	1338.7	4082.3
2.8	20.0	11	15	208.7	1183.1	3441.8	15.7	12	18	320.3	1328.7	3958.9
3.0	22.0	12	15	213.5	1228.3	3616.9	17.0	13	19	388.3	1431.8	4300.2

Note: OT = Overtime, Simulation instances = 100 for each weekday and for each risk level.

4.6.2.4 *Impact of Uncertain Patient Duration on Capacity*

In developing our capacity framework, we thus far assume deterministic patient durations in each stage. In reality, however, no two patients have the same durations even if they are classified into same patient group. Hence, we examine the impact of uncertain patient durations on capacity decision. Following an approach known as *job hedging* in the literature (Gul et al., 2011; Yellig and Mackulak, 1997), we generate patient durations as $(1 + \gamma Z)\mu$, where γ is a risk level ranging from $[0, 3]$, Z is a random variable following standard normal distribution $N(0, 1)$, and μ is the mean duration in each stage of a patient group in which a focal patient is classified.

As listed in Table 4.11, if the actual (respectively, dedicated) patient sample path strategy is selected, the total cost increase on average is 5.66% (respectively, 5.53%) as the risk level γ increases by 0.2. Also, the optimal HR capacity of the dedicated strategy is very sensitive to the duration uncertainty. For example, only 7.84% (or 1.3 units) of HRs are added if the risk level γ jumps from 0 to 3.0 for the actual strategy, whereas 90.20% (or 10.1 units) more HRs are needed for the dedicated strategy.

4.6.2.5 *Impact of Uncertain Overnight-Stay on Capacity*

We consider each patient as a potential overnight-stay patient with a certain probability, rather than assuming that the number of overnight-stay patients are given. Our ASC patient flow data includes an indicator variable predicting each patient as either overnight-stay or non-overnight-stay before surgical procedures are provided to the patients. Surprisingly, the accuracy of the indicator is only around 59%[12]. We thus generate a random variable Z following a standard normal distribution for each patient, and assign 15 hours as PACU duration if $Z < Prob(Ovngt)$, where $Prob(Ovngt) \in [5\%, 50\%]$ with 5% intervals.

As listed in Table 4.12, the 5% increase in overnight-stay probability increases on average 1.44 PACUs (respectively, 1.78) and 7.4% (respectively, 7.9%) of total cost for the actual (respectively, dedicated) patient sample path strategy.

[12]Let θ be the predicted overnight-stay and x be the actual overnight-stay indicators. In our data of 1960 patient records, $p(x|\theta) = \frac{113}{196} = 58.85\%$. $p(\theta) = \frac{192}{1960} = 9.80\%$ and $p(x) = \frac{160}{1960} = 8.16\%$.

Table 4.11: Results of Computational Experiment: Uncertain Patient Duration

	Actual Patient Sample Path						Dedicated Patient Sample Path					
Risk Lv.	#HR	#OR	#PACU	OR OT Cost	PACU OT Cost	Total Cost	#HR	#OR	#PACU	OR OT Cost	PACU OT Cost	Total Cost
0.0	16.0	8	10	117.4	931.2	2498.7	11.2	8	11	111.4	1046.7	2653.1
0.2	17.8	9	10	164.9	979.3	2719.3	12.6	9	11	161.7	1069.7	2851.4
0.4	18.9	10	11	147.0	1006.2	2898.2	14.9	11	11	143.9	1062.0	3075.9
0.6	17.5	11	11	134.2	1037.0	3041.3	16.5	12	12	98.5	1059.7	3198.2
0.8	16.1	11	11	210.8	1140.0	3220.8	16.3	12	12	164.4	1183.5	3387.9
1.0	15.0	11	11	313.0	1260.9	3443.9	15.9	12	13	223.9	1274.7	3583.6
1.2	15.3	12	11	368.5	1315.0	3678.5	16.3	13	13	256.2	1317.5	3783.7
1.4	16.8	14	12	305.9	1287.8	3883.7	15.9	13	14	347.0	1415.2	4017.2
1.6	16.3	14	12	403.0	1407.8	4100.8	15.5	13	14	463.7	1542.0	4260.7
1.8	17.2	15	13	394.1	1439.7	4293.8	16.3	14	14	522.4	1619.5	4521.9
2.0	17.0	15	13	493.5	1535.7	4489.2	16.9	15	15	574.0	1677.8	4801.9
2.2	16.7	15	13	614.1	1705.5	4779.6	16.8	15	15	687.8	1801.2	5038.9
2.4	17.6	16	14	615.4	1739.0	4984.3	18.7	17	16	631.6	1782.6	5259.2
2.6	17.4	16	14	723.8	1837.5	5191.4	20.5	19	16	581.4	1841.0	5517.4
2.8	17.4	16	14	840.3	1943.1	5413.4	21.4	20	17	551.8	1892.2	5709.0
3.0	17.3	16	14	973.2	2097.3	5700.6	21.3	20	17	658.9	2017.0	5940.8

Note: OT = Overtime, Simulation instances = 100 for each weekday and for each risk level.

Table 4.12: Results of Computational Experiment: Uncertain Overnight-Stay

.	Actual Patient Sample Path								Dedicated Patient Sample Path							
Pr(Ovngt)	#HR	#OR	#PACU	Δ#PACU	OR OT Cost	PACU OT Cost	Total Cost	ΔTotal Cost	#HR	#OR	#PACU	Δ#PACU	OR OT Cost	PACU OT Cost	Total Cost	ΔTotal Cost
5%	16	8	8	-	133.6	788.6	2282.1	-	11.178	8	9	-	109.3	874.1	2388.4	-
10%	16	8	10	2	123.7	946.5	2520.2	10.4%	11.2	8	11	2	115.5	1052.4	2662.9	11.5%
15%	16	8	11	1	134.9	1125.7	2755.7	9.3%	11.2	8	13	2	120.5	1228.2	2933.6	10.2%
20%	16	8	13	2	126.0	1276.3	2987.3	8.4%	11.2	8	14	1	144.4	1429.4	3203.8	9.2%
25%	16	8	14	1	126.6	1421.2	3177.8	6.4%	11.2	8	16	2	131.0	1589.3	3440.3	7.4%
30%	16	8	15	1	139.6	1595.9	3410.5	7.3%	12.6	9	17	1	111.7	1658.5	3660.2	6.4%
35%	16	8	17	2	135.4	1759.8	3660.2	7.3%	12.6	9	20	3	85.1	1827.9	3938.0	7.6%
40%	16	8	18	1	140.2	1930.3	3880.5	6.0%	12.6	9	22	2	104.3	2010.6	4229.8	7.4%
45%	16	8	19	1	146.6	2098.4	4100.0	5.7%	12.6	9	23	1	117.3	2180.3	4457.6	5.4%
50%	16	8	21	2	127.3	2245.6	4317.8	5.3%	12.6	9	25	2	116	2354.32	4720.32	5.9%
Average				1.44				7.4%				1.78				7.9%

Note: OT = Overtime, Simulation instances = 100 for each weekday and for each level of overnight-stay probability.

4.6.2.6 *Sensitivity Analysis on Capacity Construction Cost*

We explore the trade-off between the capacity construction cost and the overtime cost in this sensitivity analysis.

As listed in Table 4.13, the higher construction costs lead to a smaller size of capacity in each stage as expected and significantly higher total cost because of excessive overtimes. For example, if the construction costs (C_1^e, C_2^e) double from $(5.0, 3.0)$ to $(10.0, 6.0)$ while the overtime cost parameters remain as the same, the total cost increases by 147.38% (respectivley, 146.27%) for the actual (respectively, the dedicated) patient sample path strategy.

Table 4.13: Results of Computational Experiment: Sensitivity Analysis on Capacity Construction Cost:

		Actual Patient Sample Path						Dedicated Patient Sample Path					
OR Est Cost	PACU Est Cost	#HR	#OR	#PACU	OR OT Cost	PACU OT Cost	Total Cost	#HR	#OR	#PACU	OR OT Cost	PACU OT Cost	Total Cost
2.5	1.5	22	11	12	5.2	789.8	1273.7	17.7	13	13	0.5	810.3	1363.3
5.0	1.5	16	8	11	109.2	923.8	2156.8	11.2	8	12	101.8	1039.1	2275.8
5.0	3.0	16	8	10	120.4	938.1	2508.5	11.2	8	11	112.2	1061.7	2669.0
7.5	3.0	14	7	10	238.0	1068.2	3725.0	9.8	7	10	265.6	1174.8	3859.2
7.5	4.5	14	7	9	261.4	1049.4	4190.8	9.8	7	10	281.3	1187.3	4449.8
10.0	4.5	12	6	9	433.4	1211.7	5556.4	8.4	6	9	579.4	1395.5	5886.1
10.0	6.0	10	5	8	798.8	1466.8	6205.6	8.4	6	9	567.0	1385.9	6573.0

Note: OT = Overtime, Simulation instances = 100 for each weekday and for each combination of cost parameter settings.

4.7 Conclusion and Future Research Directions

Ambulatory Surgery Centers (ASCs) have transformed the outpatient experience for millions of people by offering a convenient, personalized, lower-priced alternative to hospitals. The majority of ASCs are at least partially owned by physicians, which allows for better control over scheduling. Additionally, physicians can personally guide innovative strategies for governance in responds to the dynamic environment of operations, e.g., Medicare frequently updates the list of surgeries allowable in ASCs. Extant literature suggests efficient patient scheduling decision for established facilities, but there is limited research that provides capacity planning/renovation tools under such dynamic environments, which is the focus of our study.

Benefiting from the detailed patient flow data, we provide an adaptive capacity planning framework for ASCs that have multi-stage service systems: pre-, intra-, and post-operative stages. We present the mathematical formulation to the problem based on the bin packing model to initially decide the number of operating rooms (intra-op stage) to serve planned patient demand. Afterwards, we iteratively use the hybrid flow shop model to find the efficient size of post anesthesia care unit (post-op stage) and the minimum cost flow model to find the optimal number of holding rooms (pre-op stage). In these models, the trade-off between the capacity construction cost and overtime cost guides the optimal solution. However, finding an optimal solution using the exact mathematical model is challenging. We thus propose an effective heuristic that performs well and can be applied to much larger problems. Furthermore, through computational analysis, we provide managerially relevant insights on how varying patient groups that need different amounts of service time in each stage and uncertainty in durations affect the capacity planning decisions. Future research can further implement the surgeon scheduling, which will be useful for those who already have surgeon information in the capacity planning stage.

5. CONCLUSION

My dissertation explores challenging yet encouraging payment policies and operational issues in the U.S. healthcare system. Motivated by bundled payment policies that aim to reduce practice variation, the first essay contributes by developing a precise measure of practice variation within a hospital and by examining the relationship between practice variation and hospital operational performance. From a theoretical lens of statistical process control, I empirically observe that the practice variation positively relates to the patient length of stay and the total cost per capita. Also, I find differential impacts of underuse variation in test-ordering practice on test-ordering cost and care-delivery cost, especially when the process quality and experiential quality measures are taken into account together.

Doing so enables managers and policy-makers to understand conditions on which better performance is achieved. I believe that an in-depth study on such practice variation can lead to more efficient operations of bundled payment reform models, the target opportunity of which is to reduce waste by decreasing variation in care-delivery processes. This essay documents novel evidence of how and to what extent the practice variation affects healthcare operations. Practice variation may lead to poor operational performance, and if not managed well, the excessive practice variation may diminish the benefits of quality initiatives. I hope this essay promotes research to further explicate this important practice variation metric.

Given the heterogeneity of clinical practice and performance across hospitals, the second essay devises a flexible stepwise framework for provider selection in the context of bundled payment models. Compared to the existing method that selects high ranked applicants based on a weighted average composite score (Gupta and Mehrotra, 2015), the study contributes by providing structured procedures that consider various dimensions of evaluation criteria separately and provide flexibility for participating providers to bid any combinations of care episodes they desire to join. The payer who operates this framework may also selectively decide winner groups depending on her preference/requirement of the evaluation criteria. For example, if the payer focuses on cost-efficiency (i.e., expected savings under bundled payment policy) rather than performance (i.e., care quality and productivity), the providers with high potential in saving costs will be selected. Also, the combinatorial auction model can handle geographical constraints such as the minimum number of healthcare providers to be selected in a single region. This function is beneficial as such geographical constraints are often considered by healthcare policy-makers to propagate new policies across the regions to secure patients' accessibility in an environment of limited resources. The numerical study supports the effectiveness of our selection framework compared to the weighted average method, indicating the potential suboptimal performance of existing methods (CMS, 2015a).

To perform post hoc analysis that compares the existing selection process and my framework in a more convincing manner, future research may benefit from the detailed application data of ongoing nationwide bundled payment initiative. The data includes detailed information (e.g., bundled payment care episodes, care episode lengths, discount rates, historical average payments, risk track, episode status) of each participating provider and several interesting descriptions such as the care improvement plan, the cost-saving plan, the quality measures for each care episode, and the design for gain-sharing. These detailed descriptions will enable researchers to refine my selection framework and to thoroughly compare and analyze the properties and performance of the providers selected via existing methods and the proposed selection framework.

Lastly, the third essay develops novel models that drive adaptive capacity decisions for ambulatory surgery centers (ASCs). I explicitly consider three sequential stages that are typical patient flows in ASCs: the pre-operative at holding room (HR), the intra-operative at operating room (OR), and the post-operative at post-anesthesia care unit (PACU). In general, all activities are not independent and are connected closely. Furthermore, the duration of each activity has considerable uncertainty. The interdependence of activities and uncertainties in patient-mix as well as their durations pose a significant challenge to surgical centers for managing the capacity of each activity and achieving a smooth patient flow. Under an assumption that the weekly target patient demand is defined by the ASC planners, the problem becomes how to determine the proper balance of capacity in each of the three stages and how to allocate such weekly demand to each weekday.

Considering the sequential stages with multiple rooms in each stage, ASC is modeled as a hybrid flow shop. From modeling perspectives, this study contributes by proposing an algorithm that derives the optimal size of capacity, which has not been addressed in scheduling literature. As ORs are the most expensive resource in ASCs, I first formulate a problem that provides patients' daily schedule for given fixed OR and PACU capacities. Afterward, I develop an algorithm that finds the optimal numbers of OR and PACU that minimizes the sum of overtime cost and capacity construction cost, the critical trade-off in this study. As a final step, I derive the minimum number of HRs that preserves the patient schedule determined in ORs and PACUs. Meanwhile, I provide several structural properties of the problem. Because of the computational complexity of the exact original model, I develop a heuristic that is straightforward, easy to implement, and fast enough to evaluate room capacity under uncertain business parameters such as patient-mix, service durations, overnight-stay probabilities and the relative ratio between the capacity construction cost and the overtime cost.

In contrast to the traditional top-down approach to capacity planning, my approach contributes by proposing a bottom-up strategy based on optimization methods combined with analytics that are informed by operational-level archival patient data. I expect this approach to guide ASC practitioners who are concerned with the cost and capacity implications of adding/removing specific surgical procedures in their facility. Future research may further implement surgeon scheduling, which will be useful for those who already have surgeon information in the capacity planning stage.

Using multiple research methods, including applied econometric analysis to deliver policy implications in the first essay, framework development for policy implementation in the second essay, and analytical modeling for process improvement in the third essay, my dissertation diagnoses the complexity of modern healthcare operations management and finds opportunities to improve the system. I hope my empirical findings and modeling approaches shed light on better understanding of payment reform policies, its implementation, and capacity decisions in healthcare.

REFERENCES

Abecassis, M. 2015. Bundled payment: Managing risk by managing variability. *New England Journal of Medicine Catalyst*, December 29 2015 (Accessed March 2017). URL http://catalyst.nejm.org/bundled-payment-managing-risk-by-managing-variability/.

Adan, I., J. Bekkers, N. Dellaert, J. Vissers, X. Yu. 2009. Patient mix optimisation and stochastic resource requirements: A case study in cardiothoracic surgery planning. *Health Care Management Science* **12**(2) 129.

Adida, E., H. Mamani, S. Nassiri. 2016. Bundled payment vs. fee-for-service: Impact of payment scheme on performance. *Management Science* **63**(5) 1606–1624.

Ahuja, R. K., T. L. Magnanti, J. B. Orlin. 1993. Network flows prentice hall. *Englewood Cliffs, NJ* .

Akcali, E., M. J. Cote, C. Lin. 2006. A network flow approach to optimizing hospital bed capacity decisions. *Health Care Management Science* **9**(4) 391–404.

Alexander, B. 2012. Reducing healthcare costs through appropriate test utilization. *Critical Values* **5**(2) 6–8.

Andritsos, D. A., C. S. Tang. 2014a. Introducing competition in healthcare services: The role of private care and increased patient mobility. *European Journal of Operational Research* **234**(3) 898–909.

Andritsos, D. A., C. S. Tang. 2014b. Linking process quality and resource usage: An empirical analysis. *Production and Operations Management* **23**(12) 2163–2177.

Andritsos, D. A., C. S. Tang. 2018. Incentive programs for reducing readmissions when patient care is co-produced. *Production and Operations Management* **27**(6) 999–1020.

Appleby, J., V. Raleigh, F. Frosini, G. Bevan, G. HaiYan, T. Lyscom, H. Gao. 2011. *Variations in Health Care: The Good, the Bad and the Inexplicable*. The King's Fund, London.

Arellano, M., S. Bond. 1991. Some tests of specification for panel data: Monte carlo evidence and an application to employment equations. *The Review of Economic Studies* **58**(2) 277–297.

Arellano, M., O. Bover. 1995. Another look at the instrumental variable estimation of error-components models. *Journal of Econometrics* **68**(1) 29–51.

Arrow, K., A. Auerbach, J. Bertko, S. Brownlee, L. P. Casalino, J. Cooper, F. J. Crosson, A. Enthoven, E. Falcone, R. C. Feldman. 2009. Toward a 21st-century health care system: recommendations for health care reform. *Annals of Internal Medicine* **150**(7) 493–495.

ASCA. 2011. Ambulatory Surgery Centers: A Positive Trend in Health Carem, ASC Association (Accessed June 2018). URL http://www.ascassociation.org/advancingsurgicalcare/aboutascs/industryoverview/apositivetrendinhealthcare.

Augusto, V., X. Xie, V. Perdomo. 2010. Operating theatre scheduling with patient recovery in both operating rooms and recovery beds. *Computers & Industrial Engineering* **58**(2) 231–238.

Bai, M. 2017. Optimization of surgery scheduling in multiple operating rooms with post anesthesia care unit capacity constraints: Chapter 4. surgery sequencing and scheduling in multiple ORs with PACU capacity constraints. Ph.D. thesis, Lehigh University.

Bai, M., R. H. Storer, G. L. Tonkay. 2017. A sample gradient-based algorithm for a multiple-or and pacu surgery scheduling problem. *IISE Transactions* **49**(4) 367–380.

Banker, R. D., A. Charnes, W. W. Cooper. 1984. Some models for estimating technical and scale inefficiencies in data envelopment analysis. *Management Science* **30**(9) 1078–1092.

Banker, R. D., A. Charnes, W. W. Cooper, J. Swarts, D. A. Thomas. 1989. An introduction to data envelopment analysis with some of its models and their uses. *Research in Governmental and Nonprofit Accounting* **5** 125–163.

Batun, S., B. T. Denton, T. R. Huschka, A. J. Schaefer. 2011. Operating room pooling and parallel surgery processing under uncertainty. *INFORMS Journal on Computing* **23**(2) 220–237.

Benner, M. J., M. L. Tushman. 2003. Exploitation, exploration, and process management: The productivity dilemma revisited. *Academy of Management Review* **28**(2) 238–256.

Berg, B., B. Denton, S. Erdogan, T. Rohleder, T. Huschka. 2011. Optimal booking strategies for outpatient procedure centers. *North Carolina State University Working Paper Series* .

Berg, B., B. Denton, H. Nelson, H. Balasubramanian, A. Rahman, A. Bailey, K. Lindor. 2010.

A discrete event simulation model to evaluate operational performance of a colonoscopy suite. *Medical Decision Making* **30**(3) 380–387.

Berg, B., B. T. Denton. 2012. Appointment Planning and Scheduling in Outpatient Procedure Centers. *Handbook of Healthcare System Scheduling*. Springer, 131–154.

Berlin, L. 2002. Liability for failure to order screening examinations. *American Journal of Roentgenology* **179**(6) 1401–1405.

Bhargava, H. K, A. N. Mishra. 2014. Electronic medical records and physician productivity: Evidence from panel data analysis. *Management Science* **60**(10) 2543–2562.

Blundell, R., S. Bond. 1998. Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics* **87**(1) 115–143.

Bond, S., C. Meghir. 1994. Dynamic investment models and the firm's financial policy. *The Review of Economic Studies* **61**(2) 197–222.

Bowers, J. 2013. Balancing operating theatre and bed capacity in a cardiothoracic centre. *Health Care Management Science* **16**(3) 236–244.

Brissimis, S. N., P. D. Zervopoulos. 2012. Developing a step-by-step effectiveness assessment model for customer-oriented service organizations. *European Journal of Operational Research* **223**(1) 226–233.

Cachon, G., C. Terwiesch. 2008. *Matching Supply with Demand: An Introduction to Operations Management*. McGraw-Hill, Boston.

Cardoen, B., E. Demeulemeester, J. Beliën. 2010. Operating room planning and scheduling: A literature review. *European Journal of Operational Research* **201**(3) 921–932.

Casale, A. S., R. A. Paulus, M. J. Selna, M. C. Doll, A. E. Bothe Jr., K. E. McKinley, S. A. Berry, D. E. Davis, R. J. Gilfillan, B. H. Hamory, et al. 2007. âĂIJProvenCareSMâĂİ: A provider-driven pay-for-performance program for acute episodic cardiac surgical care. *Annals of surgery* **246**(4) 613–623.

Casalino, L. P., D. Gans, R. Weber, M. Cea, A. Tuchovsky, T. F. Bishop, Y. Miranda, B. A. Frankel, K. B. Ziehler, M. M. Wong, et al. 2016. Us physician practices spend more than $15.4 billion annually to report quality measures. *Health Affairs* **35**(3) 401–406.

Cayirli, T., E. Veral. 2003. Outpatient scheduling in health care: a review of literature. *Production and Operations Management* **12**(4) 519–549.

Cayirli, T., E. Veral, H. Rosen. 2006. Designing appointment scheduling systems for ambulatory care services. *Health Care Management Science* **9**(1) 47–58.

Chandrasekaran, A., C. Senot, K. K. Boyer. 2012. Process management impact on clinical and experiential quality: Managing tensions between safe and patient-centered healthcare. *Manufacturing & Service Operations Management* **14**(4) 548–566.

Chandrashekaran, M., K. Rotte, R. Grewal. 2005. *Knowledge in error: Decoding consumer judgments with the JUMP Model*. Mahwah, NJ: Lawrence Erlbaum Associates.

Chang, J., W. Yan, H. Shao. 2004. Scheduling a two-stage no-wait hybrid flowshop with separated setup and removal times. *American Control Conference, 2004. Proceedings of the 2004*, vol. 2. IEEE, 1412–1416.

Charnes, A., W. W. Cooper, E. Rhodes. 1978. Measuring the efficiency of decision making units. *European Journal of Operational Research* **2**(6) 429–444.

Chassin, M. R. 1993. Explaining geographic variations: the enthusiasm hypothesis. *Medical Care* YS37–YS44.

Chen, L. M., A. K. Jha, S. Guterman, A. B. Ridgway, E. J. Orav, A. M. Epstein. 2010. Hospital cost of care, quality of care, and readmission rates: penny wise and pound foolish? *Archives of Internal Medicine* **170**(4) 340–346.

Cherchye, L., T. Kuosmanen, T. Post. 2000. *New Tools for Dealing with Errors-in-Variables in DEA*. Katholieke University, Center for Economic Studies, Leuven, Belgium.

Choi, K., D. Lee, D. L. Olson. 2013. Service quality and productivity in the us airline industry: A service quality-adjusted DEA model. *Service Business* 1–24.

Clancy, C. M., K. Cronin. 2005. Evidence-based decision making: global evidence, local decisions. *Health Affairs* **24**(1) 151–162.

CMS. 2010. Fiscal year 2009 quality measure reporting for 2010 payment update, Centers for Medicare & Medicaid Services (Accessed January 2017). URL http://www.cms.hhs.gov/HospitalQualityInits/downloads/HospitalRHQDAPU200808.pdf.

CMS. 2011. National Health Expenditure (NHE) fact sheet, Centers for Medicare & Medicaid Services (Accessed February 2017). URL http://www.cms.gov/NationalHealthExpendData/25_NHE_Fact_Sheet.asp.

CMS. 2013a. Bundled Payments for Care Improvement (BPCI) initiative: General Information. Centers for Medicare and Medicaid Services (Accessed Oct. 22 2016). URL http://innovation.cms.gov/initiatives/bundled-payments/index.html.

CMS. 2013b. Bundled Payments for Care Improvement Initiative: request for application, Centers for Medicare & Medicaid Services (Accessed August 6 2016).

CMS. 2014a. CMS Hospital Compare Database (Accessed Oct. 22 2016). URL https://data.medicare.gov/.

CMS. 2014b. Hospital Compare downloadable database data dictionary, Centers for Medicare & Medicaid Services (Accessed May 2018). URL https://data.medicare.gov/data/hospital-compare.

CMS. 2014c. Solicitation for applications: Acute Care Episode Demonstration by the Centers for Medicare and Medicaid Services (Accessed Aug. 12 2016). URL http://innovation.cms.gov/Files/x/ACE-Solicitation.pdf.

CMS. 2015a. CMS Bundled Payments for Care Improvement (BPCI) initiative models 2-4: Year 1 evaluation and monitoring annual report, Centers for Medicare & Medicaid Services (Accessed May 2017). URL https://innovation.cms.gov/Files/reports/BPCI-EvalRpt1.pdf.

CMS. 2015b. Risk adjustment fact sheet by the Centers for Medicare and Medicaid Services (Accessed May 2017). URL https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/PhysicianFeedbackProgram/Downloads/Risk-Adjustment-Fact-Sheet.pdf.

CMS. 2016a. CMS Bundled Payments for Care Improvement (BPCI) Initiative Models 2-4: Year 2 Evaluation and Monitoring Annual Report, Centers for Medicare & Medicaid Services (Accessed June 2016). URL https://innovation.cms.gov/Files/reports/bpci-models2-4-yr2evalrpt.

pdf.

CMS. 2016b. Design and development of the Diagnosis Related Group (DRG) by the Centers for Medicare and Medicaid Services (Accessed May 2017). URL https://www.cms.gov/ICD10Manual/version34-fullcode-cms/fullcode_cms/Design_and_development_of_the_Diagnosis_Related_Group_(DRGs)_PBL-038.pdf.

Cohen, J., P. Cohen, S. G. West, L. S. Aiken. 2003. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Lawrence Erlbaum, Mahwah, NJ.

Cooper, W. W., L. M. Seiford, J. Zhu. 2011. *Handbook on data envelopment analysis*, vol. 164. Springer.

Cooper, W. W., R. G. Thompson, R. M. Thrall. 1996. Introduction: Extensions and new developments in DEA. *Annals of Operations Research* **66**(1) 1–45.

Cramton, P., Y. Shoham, R. Steinberg. 2006. Combinatorial Auctions. MIT Press, Cambridge, MA.

Cromwell, J., D. A. Dayhoff, N. T. McCall, S. Subramanian, R. C. Freitas, R. J. Hart, et al. 1998. Medicare Participating Heart Bypass Demonstration. Executive summary. Final report. Waltham, MA: Health Economics Research, Inc.

Cromwell, J., D. A. Dayhoff, A. H. Thoumaian. 1997. Cost savings and physician responses to global bundled payments for medicare heart bypass surgery. *Health Care Financing Review* **19**(1) 41–57.

Dai, T., M. Akan, S. Tayur. 2016. Imaging room and beyond: The underlying economics behind physiciansâĂŹ test-ordering behavior in outpatient services. *Manufacturing & Service Operations Management* .

de Brantes, F., M. B. Rosenthal, M. Painter. 2009. Building a bridge from fragmentation to accountability: the Prometheus Payment model. *New England Journal of Medicine* **361**(11) 1033–1036.

Demirezen, E. M., S. Kumar, A. Sen. 2016. Sustainability of healthcare information exchanges: A game-theoretic approach. *Information Systems Research* **27**(2) 240–258.

Denoyel, V., A. Thiele, L. Alfandari. 2014. Optimal facility in-network selection for healthcare payers under reference pricing. *Working Paper* .

Denton, B., J. Viapiano, A. Vogl. 2007. Optimization of surgery sequencing and scheduling decisions under uncertainty. *Health Care Management Science* **10**(1) 13–24.

Denton, B. T., A. J. Miller, H. J. Balasubramanian, T. R. Huschka. 2010. Optimal allocation of surgery blocks to operating rooms under uncertainty. *Operations Research* **58**(4-part-1) 802–816.

Dexter, F., D. H. Penning, R. D. Traub. 2001. Statistical analysis by monte-carlo simulation of the impact of administrative and medical delays in discharge from the postanesthesia care unit on total patient care hours. *Anesthesia & Analgesia* **92**(5) 1222–1225.

Dexter, F., R. D. Traub. 2002. How to schedule elective surgical cases into specific operating rooms to maximize the efficiency of use of operating room time. *Anesthesia & Analgesia* **94**(4) 933–942.

Ding, D. X. 2014. The effect of experience, ownership and focus on productive efficiency: A longitudinal study of us hospitals. *Journal of Operations Management* **32**(1) 1–14.

Donabedian, A. 1980. *The Definition of Quality and Approaches to its Assesment*, vol. 1. Health Administration Press, Chicago, IL.

Douglas, T. J., W. Q. Judge. 2001. Total quality management implementation and competitive advantage: the role of structural control and exploration. *Academy of Management Journal* **44**(1) 158–169.

Drye, E. E., S.-L. T. Normand, Y. Wang, J. S. Ross, G. C. Schreiner, L. Han, M. Rapp, H. M. Krumholz. 2012. Comparison of hospital risk-standardized mortality rates calculated by using in-hospital and 30-day models: an observational study with implications for hospital profiling. *Annals of Internal Medicine* **156**(1_Part_1) 19–26.

Eddy, D. M. 1984. Variations in physician practice: the role of uncertainty. *Health Affairs* **3**(2) 74–89.

Elwyn, G., A. Edwards, P. Kinnersley, R. Grol. 2000. Shared decision making and the concept of equipoise: the competences of involving patients in healthcare choices. *British Journal of*

General Practice **50**(460) 892–899.

Epstein, R., L. Henríquez, J. Catalán, G. Y. Weintraub, C. Martínez. 2002. A combinational auction improves school meals in chile. *Interfaces* **32**(6) 1–14.

Estrada, J. 2007. Mean-semivariance optimization: A heuristic approach .

Evans, J. R., W. M. Lindsay. 1999. *The Management and Control of Quality*. South-Western College Pub. Cincinnati, Ohio.

Ferenc, D. P. 2013. *Understanding Hospital Billing and Coding*. Elsevier Health Sciences,.

Fisher, E. S., D. E. Wennberg, T. A. Stukel, D. J. Gottlieb, F. L. Lucas, E. L. Pinder. 2003. The implications of regional variations in medicare spending. part 1: The content, quality, and accessibility of care. *Annals of Internal Medicine* **138**(4) 273–287.

FMI. 2017. Ambulatory Surgical Centres market: Global industry analysis 2012 - 2016 and opportunity assessment 2017 - 2027. Tech. rep., Future Market Insights.

Fonarow, G. C., E. D. Peterson. 2009. Heart failure performance measures and outcomes: real or illusory gains. *Journal of the American Medical Association* **302**(7) 792–794.

Fox, J. B., F. E. Shaw. 2015. Clinical preventive services coverage and the affordable care act. *American Journal of Public Health* **105**(1) e7–e10.

Friesen, D. K. 1984. Tighter bounds for the multifit processor scheduling algorithm. *SIAM Journal on Computing* **13**(1) 170–181.

Gandhi, T. K., A. Kachalia, E. J. Thomas, A. L. Puopolo, C. Yoon, T. A. Brennan, D. M. Studdert. 2006. Missed and delayed diagnoses in the ambulatory setting: a study of closed malpractice claims. *Annals of Internal Medicine* **145**(7) 488–496.

García, J. M., S. Lozano. 2005. Production and delivery scheduling problem with time windows. *Computers & Industrial Engineering* **48**(4) 733–742.

Garvin, D. A. 1988. *Managing Quality: The Strategic and Competitive Edge*. Simon and Schuster, New York, NY.

Gerchak, Y., D. Gupta, M. Henig. 1996. Reservation planning for elective surgery under uncertain demand for emergency surgery. *Management Science* **42**(3) 321–334.

Gerrity, M. S., R. F. DeVellis, J. A. Earp. 1990. Physicians' reactions to uncertainty in patient care: a new measure and new insights. *Medical Care* 724–736.

Gicquel, C., L. Hege, M. Minoux, W. Van Canneyt. 2012. A discrete time exact solution approach for a complex hybrid flow-shop scheduling problem with limited-wait constraints. *Computers & Operations Research* **39**(3) 629–636.

Gilmore, P. C., R. E. Gomory. 1964. Sequencing a one state-variable machine: A solvable case of the traveling salesman problem. *Operations Research* **12**(5) 655–679.

Goossens, D. R., S. Onderstal, J. Pijnacker, F. C. R. Spieksma. 2014. Solids: A combinatorial auction for real estate. *Interfaces* **44**(4) 351–363.

Green, L. V. 2002. How many hospital beds? *INQUIRY: The Journal of Health Care Organization, Provision, and Financing* **39**(4) 400–412.

Green, L. V. 2012. OM forum-the vital role of operations analysis in improving healthcare delivery. *Manufacturing & Service Operations Management* **14**(4) 488–494.

Greenberg, L., D. Campion. 2006. *Efficiency in Health Care: What Does It Mean? How is it Measured? How Can it be Used for Value-Based Purchasing?*. 290-04-0001, AcademyHealth, AHRQ, 290-04-0001.

Grewal, R., M. Chandrashekaran, A. V. Citrin. 2010. Customer satisfaction heterogeneity and shareholder value. *Journal of Marketing Research* **47**(4) 612–626.

Guirchoun, S., P. Martineau, J.-C. Billaut. 2005. Total completion time minimization in a computer system with a server and two parallel processors. *Computers & Operations Research* **32**(3) 599–611.

Gul, S., B. T. Denton, J. W. Fowler, T. Huschka. 2011. Bi-criteria scheduling of surgical services for an outpatient procedure center. *Production and Operations Management* **20**(3) 406–417.

Guo, P., C. S. Tang, Y. Wang, M. Zhao. 2016. The impact of reimbursement policy on patient welfare, readmission rate and waiting time in a public healthcare system: Fee-for-service vs. bundled payment. University of California Los Angeles Working Paper. Available at http://www.anderson.ucla.edu/lib/email/FACULTY/ChrisTang_healthcaredtraft-20160708.pdf.

Gupta, D. 2007. Surgical suites' operations management. *Production and Operations Management* **16**(6) 689–700.

Gupta, D., B. Denton. 2008. Appointment scheduling in health care: Challenges and opportunities. *IIE Transactions* **40**(9) 800–819.

Gupta, D., M. Mehrotra. 2015. Bundled payments for healthcare services: Proposer selection and information sharing. *Operations Research* **63**(4) 772–788.

Gupta, D., S. J. Potthoff, et al. 2016. Matching supply and demand for hospital services. *Foundations and Trends® in Technology, Information and Operations Management* **8**(3–4) 131–274.

Ham, C. 1988. *Health Care Variations: Assessing the Evidence*. King's Fund Institute, London.

Hamilton, L. C. 2012. *Statistics with Stata: Version 12*. Cengage Learning, Boston, MA.

Hans, E. W., M. Van Houdenhoven, P. J. H. Hulshof. 2012. A framework for healthcare planning and control. *Handbook of Healthcare System Scheduling*. Springer, 303–320.

Hansen, L. P., K. J. Singleton. 1982. Generalized instrumental variables estimation of nonlinear rational expectations models. *Econometrica: Journal of the Econometric Society* 1269–1286.

Hayward, R. A. 2007. Performance measurement in search of a path. *New England Journal of Medicine* **356**(9) 951–953.

HCUP. 2012. HCUP methods series overview of key readmission measures and methods by healthcare cost and utilization project (report # 2012-04) (Accessed June 2018). URL https://www.hcup-us.ahrq.gov/reports/methods/2012_04.pdf.

HHS. 2015. Better, smarter, healthier: in historic announcement, hhs sets clear goals and timeline for shifting medicare reimbursements from volume to value, U.S. Department of Health and Human Services. *Retrieved January* **14** 2016.

Hollingsworth, B., P. J. Dawson, N. Maniadakis. 1999. Efficiency measurement of health care: a review of non-parametric methods and applications. *Health Care Management Science* **2**(3) 161–172.

Horwitz, L., C. Partovian, Z. Lin, J. Herrin, J. Grady, M. Conover, J. Montague, C. Dillaway, K. Bartczak, J. Ross. 2011. Hospital-wide (all-condition) 30-day risk-standardized readmission measure. Tech. rep., Yale New Haven Health Services Corporation Center for Outcomes Research & Evaluation.

Horwitz, L., C. Partovian, Z. Lin, J. Herrin, J. Grady, M. Conover, J. Montague, C. Dillaway, K. Bartczak, J. Ross. 2012. Hospital-wide all-cause unplanned readmission measure: final technical report. Tech. rep., Centers for Medicare and Medicaid Services.

Hsu, V. N., R. de Matta, C.-Y. Lee. 2003. Scheduling patients in an ambulatory surgical center. *Naval Research Logistics* **50**(3) 218–238.

Hussey, P. S., C. Eibner, M. S. Ridgely, E. A. McGlynn. 2009. Controlling US health care spendingâĂŤseparating promising from unpromising approaches. *New England Journal of Medicine* **361**(22) 2109–2111.

Hussey, P. S., M. S. Ridgely, M. B. Rosenthal. 2011. The Prometheus bundled payment experiment: Slow start shows problems in implementing new payment models. *Health Affairs* **30**(11) 2116–2124.

Iser, J. H., B. T. Denton, R. E. King. 2008. Heuristics for balancing operating room and post-anesthesia resources under uncertainty. *In Proceedings of the 40th Conference on Winter Simulation, Austin, TX*. IEEE, 1601–1608.

Jebali, A., A. B. H. Alouane, P. Ladet. 2006. Operating rooms scheduling. *International Journal of Production Economics* **99**(1-2) 52–62.

Jha, A. K. 2006. Measuring hospital quality: What physicians do? How patients fare? or both? *Journal of the American Medical Association* **296**(1) 95–97.

Kapu, A. N., A. P. Wheeler, B. Lee. 2014. Addition of acute care nurse practitioners to medical and surgical rapid response teams: a pilot project. *Critical Care Nurse* **34**(1) 51–59.

Kelley, K. 2007. Sample size planning for the coefficient of variation from the accuracy in parameter estimation approach. *Behavior Research Methods* **39**(4) 755–766.

Khunlertkit, A., P. Carayon. 2013. Contributions of tele–intensive care unit (Tele-ICU) technology to quality of care and patient safety. *Journal of Critical Care* **28**(3) 315–e1.

Kohn, L. T., J. M. Corrigan, M. S. Donaldson. 2000. *To Err is Human: Building a Safer Health*

System. National Academies Press, Washington DC.

Koulamas, C., G. J. Kyparisis. 2000. Asymptotically optimal linear time algorithms for two-stage and three-stage flexible flow shops. *Naval Research Logistics* **47**(3) 259–268.

Krackhardt, D., L. W. Porter. 1986. The snowball effect: Turnover embedded in communication networks. *Journal of Applied Psychology* **71**(1) 50.

Kuhnen, C. M., A. Niessen. 2012. Public opinion and executive compensation. *Management Science* **58**(7) 1249–1272.

Kwasnica, A. M., J. O. Ledyard, D. Porter, C. DeMartini. 2005. A new and improved design for multiobject iterative auctions. *Management Science* **51**(3) 419–434.

Lake, E. T., C. R. Friese. 2006. Variations in nursing practice environments: relation to staffing and hospital characteristics. *Nursing Research* **55**(1) 1–9.

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

Leape, L. L. 1994. Error in medicine. *Journal of the American Medical Association* **272**(23) 1851–1857.

Leavitt, M. O. 2006. Report to Congress: Physician group practice demonstration, first evaluation report. United States Department of Health and Human Services, Washington D.C.

Lee, H. L., V. Padmanabhan, S. Whang. 1997. The bullwhip effect in supply chains. *Sloan Management Review* **38**(3) 93.

Lee, S., Y. Yih. 2012. Surgery scheduling of multiple operating rooms under uncertainty and resource constraints of post-anesthesia care units. *IIE Annual Conference. Proceedings*. Institute of Industrial and Systems Engineers, 1.

Lee, S., Y. Yih. 2014. Reducing patient-flow delays in surgical suites through determining start-times of surgical cases. *European Journal of Operational Research* **238**(2) 620–629.

Lee, S. J., G. R. Heim, C. Sriskandarajah, Y. Zhu. 2018. Outpatient appointment block scheduling under patient heterogeneity and patient no-shows. *Production and Operations Management*

27(1) 28–48.

Lemay, B., A. Cohn, M. Epelman, S. Gorga. 2017. New methods for resolving conflicting requests with examples from medical residency scheduling. *Production and Operations Management* **26**(9) 1778–1793.

Li, L. X., W. C. Benton. 1996. Performance measurement criteria in health care organizations: review and future research directions. *European Journal of Operational Research* **93**(3) 449–468.

Lilford, R. J., C. A. Brown, J. Nicholl. 2007. Use of process measures to monitor the quality of clinical practice. *British Medical Journal* **335**(7621) 648.

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

Lindenberg, S. 1990. Homo socio-oeconomicus: The emergence of a general model of man in the social sciences. *Journal of Institutional and Theoretical Economics* **146**(4) 727–748.

Liu, N., V.-A. Truong, X. Wang, B. Anderson. 2019. Integrated scheduling and capacity planning with considerations for patients' length-of-stays. *Production and Operations Management* doi: 10.1111/poms.13012.Forthcoming.

Liu, N., S. Ziya, V. G. Kulkarni. 2010. Dynamic scheduling of outpatient appointments under patient no-shows and cancellations. *Manufacturing & Service Operations Management* **12**(2) 347–364.

Liu, Y., I. Karimi. 2008. Scheduling multistage batch plants with parallel units and no interstage storage. *Computers & Chemical Engineering* **32**(4) 671–693.

Lorenz, E. 2000. The butterfly effect. *World Scientific Series on Nonlinear Science Series A* **39** 91–94.

Luo, X., S. Raithel, M. A. Wiles. 2013. The impact of brand rating dispersion on firm value. *Journal of Marketing Research* **50**(3) 399–415.

Mak, H.-Y., Y. Rong, J. Zhang. 2014. Sequencing appointments for service systems using inventory approximations. *Manufacturing & Service Operations Management* **16**(2) 251–262.

Mancilla, C., R. Storer. 2012. A sample average approximation approach to stochastic appointment sequencing and scheduling. *IIE Transactions* **44**(8) 655–670.

March, J. G. 1991. Exploration and exploitation in organizational learning. *Organization Science* **2**(1) 71–87.

Marcon, E., F. Dexter. 2006. Impact of surgical sequencing on post anesthesia care unit staffing. *Health Care Management Science* **9**(1) 87–98.

Mascis, A., D. Pacciarelli. 2002. Job-shop scheduling with blocking and no-wait constraints. *European Journal of Operational Research* **143**(3) 498–517.

May, J. H., W. E. Spangler, D. P. Strum, L. G. Vargas. 2011. The surgical scheduling problem: Current research and future opportunities. *Production and Operations Management* **20**(3) 392–405.

McCabe, B., V. Tran, J. Ramani. 2005. Construction prequalification using data envelopment analysis. *Canadian Journal of Civil Engineering* **32**(1) 183–193.

MedPAC. 2017. Medicare payment advisory commission 2017 report to congress (Accessed Aug. 05 2018). URL http://www.medpac.gov/documents.

Melnick, G. A., K. Fonkych. 2008. Hospital pricing and the uninsured: do the uninsured pay higher prices? *Health Affairs* **27**(2) w116–w122.

Meltzer, D. O., J. W. Chung. 2014. The population value of quality indicator reporting: a framework for prioritizing health care performance measures. *Health affairs* **33**(1) 132–139.

Miller, D. C., C. Gust, J. B. Dimick, N. Birkmeyer, J. Skinner, J. D. Birkmeyer. 2011. Large variations in Medicare payments for surgery highlight savings potential from bundled payment programs. *Health Affairs* **30**(11) 2107–2115.

Miller, H. D. 2009. From volume to value: Better ways to pay for health care. *Health Affairs* **28**(5) 1418–1428.

Min, D., Y. Yih. 2010. Scheduling elective surgery under uncertainty and downstream capacity constraints. *European Journal of Operational Research* **206**(3) 642–652.

Mitchell, K. 2014. Measuring what matters: CMS introduces a new methodology for "topped out" measures (Accessed June 2016). *Morning Consult*, October 23 2014 (Accessed June 2018). URL http://morningconsult.com/opinions/measuring-matters-cms-introduces-new-methodology-topped-measures/.

Mossialos, E., M. Wenzl, R. Osborn, C. Anderson. 2016. *2015 International Profiles of Health Care Systems*. The Commonwealth Fund, New York, NY.

Muthuraman, K., M. Lawley. 2008. A stochastic overbooking model for outpatient clinical scheduling with no-shows. *IIE Transactions* **40**(9) 820–837.

Nair, A., M. Nicolae, R. Narasimhan. 2013. Examining the impact of clinical quality and clinical flexibility on cardiology unit performance: Does experiential quality act as a specialized complementary asset? *Journal of Operations Management* **31**(7) 505–522.

Narahari, Y., P. Dayama. 2005. Combinatorial auctions for electronic business. *Sadhana - Indian Academy Proceedings in Engineering Sciences*, vol. 30. Indian Academy of Sciences, 179–211.

Narayan, V., V. Kadiyali. 2015. Repeated interactions and improved outcomes: An empirical analysis of movie production in the united states. *Management Science* **62**(2) 591–607.

NASEM. 2016. *Improving diagnosis in health care, National Academies of Sciences, Engineering, and Medicine*. National Academies Press.

Nutting, P. A. 1994. Competing demands of primary care: a model for the delivery of clinical preventive services. *The Journal of Family Practice* **38**(0) 167.

Oakland, J. S. 2007. *Statistical Process Control*. Routledge, New York, NY.

OECD. 2016. OECD Health statistics 2016 (Accessed May 2018). URL https://innovation.cms.gov/Files/x/ACE-Solicitation.pdf.

Olivares, M., G. Y. Weintraub, R. Epstein, D. Yung. 2012. Combinatorial auctions for procurement: An empirical study of the chilean school meals auction. *Management Science* **58**(8) 1458–1481.

Ozcan, Y. A. 2008. *Health Care Benchmarking and Performance Evaluation: An Assessment using Data Envelopment Analysis (DEA)*. Springer, Berlin.

Ozcan, Y. A., J. R. Lynch. 1991. Rural hospital closures: An inquiry into efficiency. *Advances in*

Health Economics and Health Services Research **13** 205–224.

Pang, M.-S., A. Tafti, M. S. Krishnan. 2016. Do cio it budgets explain bigger or smaller governments? theory and evidence from us state governments. *Management Science* **62**(4) 1020–1041.

Patrick, J., M. L. Puterman, M. Queyranne. 2008. Dynamic multipriority patient scheduling for a diagnostic resource. *Operations Research* **56**(6) 1507–1525.

Paulus, R. A., K. Davis, G. D. Steele. 2008. Continuous innovation in health care: Implications of the Geisinger experience. *Health Affairs* **27**(5) 1235–1245.

Perdikaki, O., S. Kesavan, J. M. Swaminathan. 2012. Effect of traffic on sales and conversion rates of retail stores. *Manufacturing & Service Operations Management* **14**(1) 145–162.

Pham, D.-N., A. Klinkert. 2008. Surgical case scheduling as a generalized job shop scheduling problem. *European Journal of Operational Research* **185**(3) 1011–1025.

Pinedo, M. L. 2016. *Scheduling: Theory, Algorithms, and Systems*. Springer.

Porter, M. E., T. H. Lee. 2013. The strategy that will fix health care. *Harvard Business Review* **91**(12) 24.

Powell, T. C. 1995. Total quality management as competitive advantage: a review and empirical study. *Strategic Management Journal* **16**(1) 15–37.

Price, C., B. Golden, M. Harrington, R. Konewko, E. Wasil, W. Herring. 2011. Reducing boarding in a post-anesthesia care unit. *Production and Operations Management* **20**(3) 431–441.

Pulido, R., A. M. Aguirre, M. Ortega-Mier, Á. García-Sánchez, C. A. Méndez. 2014. Managing daily surgery schedules in a teaching hospital: a mixed-integer optimization approach. *BMC Health Services Research* **14**(1) 464.

Rego, L. L., N. A. Morgan, C. Fornell. 2013. Reexamining the market share–customer satisfaction relationship. *Journal of Marketing* **77**(5) 1–20.

Ribas, I., R. Leisten, J. M. Framiñan. 2010. Review and classification of hybrid flow shop scheduling problems from a production system and a solutions procedure perspective. *Computers & Operations Research* **37**(8) 1439–1454.

Robinson, L. W., R. R. Chen. 2003. Scheduling doctors' appointments: optimal and empirically-based heuristic policies. *IIE Transactions* **35**(3) 295–307.

Rohatgi, V. K. 2011. Semi-variance in finance. *International Encyclopedia of Statistical Science*. Springer, 1297–1298.

Roodman, D. 2006. How to do xtabond2: An introduction to difference and system GMM in stata. *Stata Journal* **9**(1) 86–136.

Rosenthal, E. 2013. As hospital prices soar, a stitch tops $500. *New York Times*, December 2 2013 (Accessed August 2017). URL https://www.nytimes.com/2013/12/03/health/as-hospital-costs-soar-single-stitch-tops-500.html?pagewanted=all&_r=0.

Rubin, H. R., P. Pronovost, G. B. Diette. 2001. The advantages and disadvantages of process-based measures of health care quality. *International Journal for Quality in Health Care* **13**(6) 469–474.

Ruiz, R., J. A. Vázquez-Rodríguez. 2010. The hybrid flow shop scheduling problem. *European Journal of Operational Research* **205**(1) 1–18.

Sadeghi, S., A. Barzi, M. M. Shabot, O. Mikhail. 2012. *Integrating Quality and Strategy in Health Care Organizations*. Jones & Bartlett Publishers, Burlington, MA.

Salzarulo, P. A., K. M. Bretthauer, M. J. Côté, K. L. Schultz. 2011. The impact of variability and patient information on health care system performance. *Production and Operations Management* **20**(6) 848–859.

Satin, D. J., J. Miles. 2009. Performance-based bundled payments: potential benefits and burdens. *Minnesota Medicine* **92**(10) 33–35.

Senot, C., A. Chandrasekaran, P. T. Ward, A. L. Tucker, S. D. Moffatt-Bruce. 2015. The impact of combining conformance and experiential quality on hospitalsâĂŹ readmissions and cost performance. *Management Science* **62**(3) 829–848.

Sherman, H. D., J. Zhu. 2006. Benchmarking with quality-adjusted DEA (Q-DEA) to seek lower-cost high-quality service: Evidence from a US bank application. *Annals of Operations Research* **145**(1) 301–319.

Shi, L., D. A. Singh. 2012. *Essentials of the US Health Care System*. Jones & Bartlett Publishers, Burlington, MA.

Shimshak, D. G., M. L. Lenard, R. K. Klimberg. 2009. Incorporating quality into data envelopment analysis of nursing home performance: A case study. *Omega* **37**(3) 672–685.

Shwartz, M., A. B. Cohen, J. D. Restuccia, Z. J. Ren, A. Labonte, C. Theokary, R. Kang, J. Horwitt. 2011. How well can we identify the high-performing hospital? *Medical Care Research and Review* **68**(3) 290–310.

Sipkoff, M. 2003. 9 ways to reduce unwarranted variation. *Managed Care* **12**(11) 20–4.

Smolowitz, J., E. Speakman, D. Wojnar, E.-M. Whelan, S. Ulrich, C. Hayes, L. Wood. 2015. Role of the registered nurse in primary health care: Meeting health care needs in the 21st century. *Nursing Outlook* **63**(2) 130–136.

Sriskandarajah, C. 1993. Performance of scheduling algorithms for no-wait flowshops with parallel machines. *European Journal of Operational Research* **70**(3) 365–378.

Sriskandarajah, C., P. Ladet. 1986. Some no-wait shops scheduling problems: complexity aspect. *European Journal of Operational Research* **24**(3) 424–438.

Stano, M. 1993. Evaluating the policy role of the small area variations and physician practice style hypotheses. *Health Policy* **24**(1) 9–17.

Theokary, C., Z. J. Ren. 2011. An empirical study of the relations between hospital volume, teaching status, and service quality. *Production and Operations Management* **20**(3) 303–318.

Thornton, H. W., J. L. Hunsucker. 2004. A new heuristic for minimal makespan in flow shops with multiple processors and no intermediate storage. *European Journal of Operational Research* **152**(1) 96–114.

Tiwari, V., W. S. Sandberg. 2016. Perioperative bed capacity planning guided by theory of constraints. *Winter Simulation Conference (WSC), 2016*. IEEE, 1894–1903.

Toussaint, J. 2009. Writing the new playbook for us health care: lessons from wisconsin. *Health Affairs* **28**(5) 1343–1350.

Tsai, T. C., K. E. Joynt, R. C. Wild, E. J. Orav, A. K. Jha. 2015. Medicare's bundled payment initiative: most hospitals are focused on a few high-volume conditions. *Health Affairs* **34**(3)

371–380.

Vanberkel, P. T., M. E. Zonderland, R. J. Boucherie, E. W. Hans. 2012. Leaving a mark on health-care delivery with operations analysis. *Manufacturing & Service Operations Management* **14**(4) 508–511.

VMG-Health. 2017. Intellimarker multi-specialty ASC study (Accessed August 2018). URL https://vmghealth.com/wp-content/uploads/2018/01/VMG-Health-Intellimarker-Multi-Specialty-ASC-Study-2017.pdf.

Vujcic, B., I. Jemric. 2001. Efficiency of banks in transition: A DEA approach, Croatian National Bank.

Wang, S., M. Liu. 2013. A genetic algorithm for two-stage no-wait hybrid flow shop scheduling problem. *Computers & Operations Research* **40**(4) 1064–1075.

Wang, S., M. Liu, C. Chu. 2015. A branch-and-bound algorithm for two-stage no-wait hybrid flow-shop scheduling. *International Journal of Production Research* **53**(4) 1143–1167.

Wang, Z., W. Xing, F. Bai. 2005. No-wait flexible flowshop scheduling with no-idle machines. *Operations Research Letters* **33**(6) 609–614.

Wennberg, J., A. Gittelsohn. 1982. Variations in medical care among small areas. *Scientific American* **246**(4) 120–134.

Wennberg, J. E. 2002. Unwarranted variations in healthcare delivery: implications for academic medical centres. *British Medical Journal* **325**(7370) 961.

Wennberg, J. E., E. S. Fisher, J. S. Skinner. 2002. Geography and the debate over medicare reform. *Health Affairs* **21**(2) 10–10.

Wennberg, J. E., A. M. Gittelsohn. 1973. Small area variations in health care delivery. American Association for the Advancement of Science , Washington, DC.

Westert, G. P., P. P. Groenewegen. 1999. Medical practice variations: changing the theoretical approach. *Scandinavian Journal of Public Health* **27**(3) 173–180.

Wheeler, D. J., D. S. Chambers, et al. 1992. *Understanding statistical process control*. SPC press.

Wintoki, M. B., J. S. Linck, J. M. Netter. 2012. Endogeneity and the dynamics of internal corporate governance. *Journal of Financial Economics* **105**(3) 581–606.

Xie, J., W. Xing, Z. Liu, J. Dong. 2004. Minimum deviation algorithm for two-stageno-wait flowshops with parallel machines. *Computers & Mathematics with Applications* **47**(12) 1857–1863.

Yang, J. 2010. A new complexity proof for the two-stage hybrid flow shop scheduling problem with dedicated machines. *International Journal of Production Research* **48**(5) 1531–1538.

Yellig, E. J., G. T. Mackulak. 1997. Robust deterministic scheduling in stochastic environments: The method of capacity hedge points. *International Journal of Production Research* **35**(2) 369–379.

Zervopoulos, P., T. Palaskas. 2011. Applying quality-driven, efficiency-adjusted DEA (QE-DEA) in the pursuit of high-efficiency–high-quality service units: An input-oriented approach. *IMA Journal of Management Mathematics* **22**(4) 401–417.

Zhi, M., E. L. Ding, J. Theisen-Toupal, J. Whelan, R. Arnaout. 2013. The landscape of inappropriate laboratory testing: a 15-year meta-analysis. *PloS one* **8**(11) e78962.

APPENDIX A

SUPPLEMENT TO CHAPTER 2

A.1 Table of Acronyms

Table A.1 contains a list of acronyms used in Chapter 2.

Table A.1: Table of Acronyms in Chapter 2

Acronym	Meaning
ACR	American College of Radiology
AHA	American Heart Association
AHRQ	Association for Healthcare Research and Quality
AMI	Acute Myocardial Infarction
BPCI	Bundled Payments for Care Improvement
CBSA	Core-Based Statistical Area
CDM	Charge Description Master
CMI	Case Mix Index
CMS	Centers for Medicare & Medicaid Services
CV	Coefficient of Variation
EQ	Experiential Quality
FFS	Fee For Service
GMM	Generalized Method of Moments
HCAHPS	Hospital Consumer Assessment of Healthcare Providers and Systems
HCUP	Healthcare Cost and Utilization Project
HF	Heart Failure
HHS	Department of Health and Human Services
HIMSS	Healthcare Information and Management Systems Society
HIT	Health Information Technologies
HQA	Hospital Quality Alliance
ICU	Intensive Care Unit
IV	Instrumental Variable
LOS	Length-of-Stay
MS-DRG	Medicare Severity - Diagnosis Related Group
MSPB	Medicare Spending Per Beneficiary
PN	Pneumonia
PQ	Process Quality
SID	State Inpatient Discharge
SPC	Statistical Processing Control
TQM	Total Quality Management
WACV	Weighted Average Coefficient of Variation

A.2 Variable Descriptions

The variables used in this study are summarized in Table A.2.

Table A.2: Variable Descriptions

Variable	Level	Description	Source
		Dependent Variables	
TotalLOS	Hos-Cond[a]	Logarithm of risk-adjusted average total LOS (Sum of LOS per discharge if a focal patient is readmitted within 30 days).	HCUP
TotalCost	Hos-Cond	Logarithm of risk- and inflation-adjusted average total cost (Sum of cost per discharge if a focal patient is readmitted within 30 days). Note: TotalCost = TestCost + CareCost.	HCUP
TestCost	Hos-Cond	Logarithm of risk- and inflation–adjusted average cost related to ordering laboratory/radiology test (Sum of related costs if a focal patient is readmitted within 30 days).	HCUP
CareCost	Hos-Cond	Logarithm of risk- and inflation–adjusted average cost of delivering care (Sum of related costs if a focal patient is readmitted within 30 days).	HCUP
		Independent and Moderator Variables	
$\text{WACVG}^{T}_{\text{all}}$	Hos-Cond	Practice variation during a patient care-episode (patient cohort defined by MS-DRG code). Target Standard $T \in \{h, c, s\}$ where h=Hospital-average, c=County-average, and s=State/CBSA-average.	HCUP
$\text{WACVG}^{T}_{\text{under}}$	Hos-Cond	Underuse practice variation in ordering laboratory/radiology tests (patient cohort defined by MS-DRG code). Target Standard $T \in \{h, c, s\}$	HCUP
$\text{WACVG}^{T}_{\text{over}}$	Hos-Cond	Overuse practice variation in ordering laboratory/radiology tests (patient cohort defined by MS-DRG code). Target Standard $T \in \{h, c, s\}$	HCUP
$\text{WACVD}^{T}_{\text{all}}$	Hos-Cond	Practice variation during a patient care-episode (patient cohort defined by principal diagnosis code). Target Standard $T \in \{h, c, s\}$ where h=Hospital-average, c=County-average, and s=State/CBSA-average.	HCUP
$\text{WACVD}^{T}_{\text{under}}$	Hos-Cond	Underuse practice variation in ordering laboratory/radiology tests (patient cohort defined by principal diagnosis code). Target Standard $T \in \{h, c, s\}$	HCUP
$\text{WACVD}^{T}_{\text{over}}$	Hos-Cond	Underuse practice variation in ordering laboratory/radiology tests (patient cohort defined by principal diagnosis code). Target Standard $T \in \{h, c, s\}$	HCUP
ProcQual	Hos-Cond	How well each hospital adhere to standardized guidelines endorsed as best practices	CMS HC[b]
ExpeQual	Hospital	Patient perceptions of the level of interaction with their caregivers during their hospital stay (HCAHPS[c] survey)	CMS HC
		Control Variables	
Bed Size	Hospital	ln(Number of Beds) in a hospital.	CMS IF[d]
CMI	Hospital	Case Mix Index reflects the diversity, clinical complexity, and the needs for resources in the population of all the patients in the hospital.	CMS IF
Teaching	Hospital	Teaching intensity defined as resident-to-bed ratios.	CMS IF
Wage Index	Hospital	Index intended to measure differences in hospital wage rates across labor markets.	CMS IF
Outlier Adj	Hospital	CMS operating outlier adjustment factor, which reflects unusually costly cases treated by the focal hospital.	CMS IF
OPDSH Adj	Hospital	CMS operating disproportionate share hospital payment adjustment factor, which reflects the hospital's propensity to treat uninsured and Medicaid patients who need more resources in general.	CMS IF
Governmental	Hospital	1 = Government sponsored, 0 = Private.	CMS HC
Nonprofit	Hospital	1 = Nonprofit, 0 = For profit.	CMS HC
Urban	Hospital	1 = Located in urban area, 0 = rural area.	CMS CR[e]
		Additional Variables for Risk-Adjustment of Dependent Variables	
Agebin	Patient	Age bins with each representing a 5-year period (e.g., Age 50 to 54).	HCUP
Gender	Patient	1 = Female, 0 = Male.	HCUP
Ethnicity	Patient	White, Black, Hispanic, Asian, etc.	HCUP
ComorIndex	Patient	The Elixhauser Comorbidity Index: 29 comorbidities (e.g., AIDS, depression).	HCUP

Note: LOS, Cost, WACV related measures, ProcQual, and ExpeQual are variables that are calculated using data from the corresponding sources.

[a] Hos-Cond: Hospital-Condition level. It can be aggregated to Hospital level; [b] CMS HC: CMS Hospital Compare Database;

[c] HCAHPS: Hospital Consumer Assessment of Healthcare Providers and Systems; [d] CMS IF: CMS Impact Files; [e] CMS CR: CMS Cost Reports.

A.3 Decision on Revisits from HCUP SID Data

In the HCUP SID Data from NY and FL states, a metric called $visitlink$ is provided that enables us to track whether a focal patient revisited the hospital in a fiscal year. We illustrate this in the four cases below to show how we determine total LOS and total cost. See Table A.3.

Table A.3: Example of HCUP data structure

Case	Visitlink	Revisit ID	Days to Event	LOS	DRG	Total Charges	Days Btwn Admission	Readmission Indicator
A	22	1	17860	2	293	2288	.	0
	22	2	17884	2	292	15570	22	1
...								
B	143	0	17806	5	291	55467	0	0
...								
C	426	1	15098	3	292	22210	.	0
	426	2	15127	2	293	6487	27	1
	426	3	15138	3	293	14286	8	1
...								
D	1271	1	16033	2	292	10184	0	0
	1271	2	16053	8	291	26568	12	1
	1271	3	16165	4	292	14916	108	0
	1271	4	16246	2	291	17937	79	0

Case A: This patient (visitlink ID=22) is readmitted in 22 days after the initial discharge. We consider these two discharge records as a same claim. Thus, Total LOS = 4, Initial LOS = 2, Readmission LOS = 2, and Total Charge = 5588+15570 = 21158.

Case B: Most frequent case. Single visit no readmission throughout a given year and state.

Case C: This patient (visitlink ID=426) is readmitted two times, i.e., 27 days after the initial discharge and 8 days after the second discharge. Because the days between two consecutive admission are less than 30 days, we still consider the three discharge records as a single claim. Total LOS = 8, Initial LOS = 3, Readmission LOS = 5, and Total Charge = sum of three Total Charges.

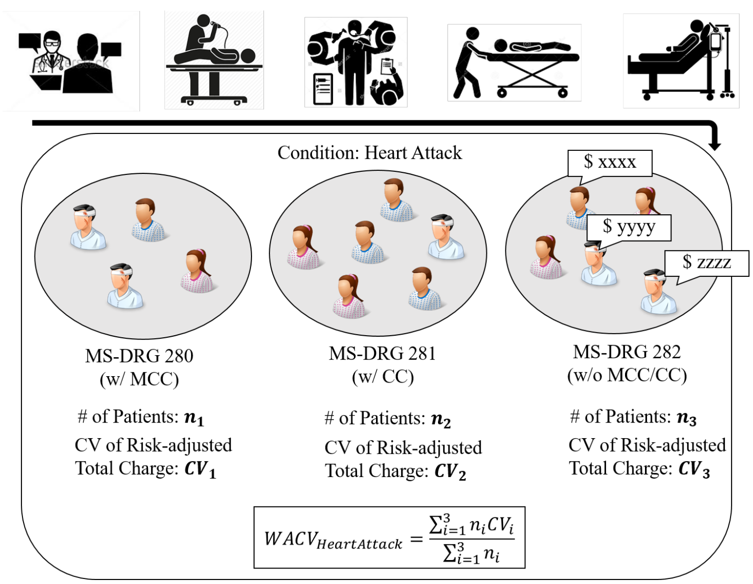
Case D: Most complicated case. This patient (visitlink 1271) is readmitted four times in a year. Based on the 30-day readmission rule adopted by CMS, we consider the first two discharge records as a single claim, and the other two records are separate claims since the days between the admissions are greater than 30 days. Thus, there are three independent claims for this patient.

We follow CMS guidelines to distinguish between unplanned and planned readmissions since some readmissions are medically necessary and thus unavoidable (Horwitz et al., 2012). For example, (1) some types of care are always considered planned (obstetrical delivery, transplant surgery, maintenance chemotherapy, rehabilitation); (2) otherwise, a planned readmission is defined as a non-acute readmission for a scheduled procedure; and (3) admissions for acute illness or for complications of care are never planned (HCUP, 2012).

A.4 Measurement of Practice Variation: An Illustration

Figure A.1 illustrates how the measure of practice variation is constructed after patients are classified into one of the MS-DRG codes. We calculate the coefficient of variation (CV) of total charges for patients within the same MS-DRG code, and then obtain a weighted-average value at the medical condition- or hospital-level, with the weight as the number of patients in each MS-DRG code.

Figure A.1: Illustration: Measurement of Practice Variation



A.5 Summary Statistics, Admission Types, and Race of Patient Sample by Condition

A.5.1 Summary Statistics of Patient Sample

Table A.4 lists summary statistics by condition for the patient sample used in our analysis. Within each condition, patients with comorbidities tend to have a greater number of diagnoses and procedures with longer LOS and higher total charges.

Table A.4: Patient Sample Summary Statistics by Condition

Condition	AMI											
MS-DRG	280 (w/ MCC)				281 (w/ CC)				282 (w/o MCC/CC)			
	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max
Age	79.65	10.68	20	112	79.06	10.55	21	107	77.87	10.54	23	107
Female	0.52	0.50	0	1	0.54	0.50	0	1	0.52	0.50	0	1
# Comorbidities	3.65	1.71	0	12	3.07	1.58	0	11	2.20	1.32	0	9
# Diagnoses	15.18	5.24	2	31	12.49	4.50	2	31	8.82	3.75	1	31
# Procedures	2.22	2.43	0	19	1.76	1.93	0	15	1.52	1.77	0	14
LOS (in Days)	6.81	5.52	0	170	4.29	3.19	0	87	2.66	2.08	0	59
Total Charges ($)	49,060.64	40,372.08	673	358,844	32,685.22	22,359.31	1,077	189,394	23,708.13	16,587.78	282	150,636
N	93,639				52,758				38,211			
Condition	HF											
MS-DRG	291 (w/ MCC)				292 (w/ CC)				293 (w/o MCC/CC)			
	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max
Age	78.03	11.86	18	113	78.68	11.16	20	111	79.90	10.36	19	110
Female	0.52	0.50	0	1	0.53	0.50	0	1	0.54	0.50	0	1
# Comorbidities	3.94	1.71	0	13	3.49	1.64	0	11	2.63	1.44	0	10
# Diagnoses	15.14	5.22	2	31	13.54	4.56	2	31	10.52	3.94	1	31
# Procedures	1.29	1.78	0	19	0.74	1.31	0	15	0.47	1.02	0	13
LOS (in Days)	6.79	5.44	0	186	4.92	3.57	0	149	3.34	2.31	0	79
Total Charges ($)	41,174.71	34,417.86	703	301,674	27,331.79	20,148.72	234	205,894	18,688.38	12,445.40	144	147,858
N	210,558				217,342				117,654			
Condition	PN											
MS-DRG	193 (w/ MCC)				194 (w/ CC)				195 (w/o MCC/CC)			
	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max
Age	76.79	13.07	18	113	77.13	12.21	18	120	77.64	11.95	18	111
Female	0.52	0.50	0	1	0.55	0.50	0	1	0.53	0.50	0	1
# Comorbidities	4.54	1.89	0	13	3.68	1.73	0	13	2.64	1.51	0	11
# Diagnoses	14.08	5.09	2	31	11.45	4.44	2	31	8.38	3.72	1	31
# Procedures	1.19	1.69	0	17	0.74	1.36	0	15	0.48	1.08	0	14
LOS (in Days)	7.15	5.20	0	313	5.17	3.40	0	103	3.68	2.33	0	66
Total Charges ($)	41,617.19	31,143.99	440	239,904	27,687.79	19,492.20	770	147,841	19,636.97	12,618.53	590	95,451
N	112,524				170,278				81,147			

We also observe from this table that the standard deviation of the total charges is very large, even for a single DRG code classified within either AMI, HF, or PN. This phenomenon is also common for other medical conditions. For example, Rosenthal (2013) describes two patients having the same deep lacerations, but who ended up with very different charges.

A.5.2 Admission Types and Race of Patient Sample

Table A.5 lists admission types and race by condition for the patient sample used in our analysis. For all three medical conditions, we observe that a majority of patients are classified as emergency or urgent.

A.6 Measures and Summary Statistics for Process Quality & Experiential Quality

Table A.6 lists the yearly individual measures for process quality and experiential quality, and their summary statistics, from 2007 to 2013. We observe that the list of process quality measures varies somewhat across years.

A.7 Correlation Table for Key Variables

Table A.7 reports correlations between lagged key variables and practice variation measures. We observe that lagged operational performance variables and control variables are correlated with our main explanatory variables, providing support for the use of dynamic panel GMM. Additionally, Table A.8 reports correlations between alternative practice variation measures.

Table A.5: Admission Types and Race of the Patient Sample by Condition

Condition		AMI					
MS-DRG		280 (w/ MCC)		281 (w/ CC)		282 (w/o MCC/CC)	
		Freq.	Percent	Freq.	Percent	Freq.	Percent
Admission Type	Emergency	83,315	88.97	45,916	87.03	33,179	86.83
	Urgent	8,352	8.92	5,411	10.26	3,975	10.40
	Elective	1,889	2.02	1,396	2.65	1,030	2.70
	Others	61	0.06	23	0.04	20	0.05
	Missing	22	0.02	12	0.02	7	0.02
	Total	93,639	100.00	52,758	100.00	38,211	100.00
Race	White	69,425	74.14	39,632	75.12	28,294	74.05
	Black	9,863	10.53	5,119	9.70	3,312	8.67
	Hispanic	8,953	9.56	5,178	9.81	4,244	11.11
	Asian	1,064	1.14	447	0.85	387	1.01
	Native American	196	0.21	72	0.14	387	1.01
	Other	3,759	4.01	2,081	3.94	68	0.18
	Missing	379	0.40	229	0.43	204	0.53
	Total	93,639	100.00	52,758	100.00	38,211	100.00
Condition		HF					
MS-DRG		291 (w/ MCC)		292 (w/ CC)		293 (w/o MCC/CC)	
		Freq.	Percent	Freq.	Percent	Freq.	Percent
Admission Type	Emergency	189,100	89.81	190,523	87.66	103,914	88.32
	Urgent	16,078	7.64	19,534	8.99	9,974	8.48
	Elective	5,278	2.51	7,201	3.31	3,710	3.15
	Others	74	0.03	59	0.03	43	0.03
	Missing	28	0.01	25	0.01	13	0.01
	Total	210,558	100.00	217,342	100.00	117,654	100.00
Race	White	144,002	68.39	152,398	70.12	80,658	68.56
	Black	33,010	15.68	33,214	15.28	18,069	15.36
	Hispanic	23,113	10.98	21,633	9.95	12,979	11.03
	Asian	2,163	1.03	1,697	0.78	990	0.84
	Native American	423	0.20	357	0.16	260	0.22
	Other	7,215	3.43	7,456	3.43	4,349	3.70
	Missing	632	0.30	587	0.27	349	0.30
	Total	210,558	100.00	217,342	100.00	117,654	100.00
Condition		PN					
MS-DRG		193 (w/ MCC)		194 (w/ CC)		195 (w/o MCC/CC)	
		Freq.	Percent	Freq.	Percent	Freq.	Percent
Admission Type	Emergency	101,453	90.16	150,432	88.34	72,163	88.93
	Urgent	8,355	7.43	14,668	8.61	6,628	8.17
	Elective	2,675	2.38	5,100	3.00	2,296	2.83
	Others	29	0.03	43	0.02	43	0.05
	Missing	12	0.01	35	0.02	17	0.02
	Total	112,524	100.00	170,278	100.00	81,147	100.00
Race	White	82,919	73.69	130,387	76.57	60,872	75.01
	Black	13,039	11.59	14,577	8.56	7,220	8.90
	Hispanic	11,111	9.87	17,012	9.99	8,853	10.91
	Asian	1,368	1.22	2,108	1.24	1,058	1.30
	Native American	228	0.20	306	0.18	183	0.23
	Other	3,461	3.08	5,312	3.12	2,642	3.26
	Missing	398	0.35	576	0.34	319	0.39
	Total	112,524	100.00	170,278	100.00	81,147	100.00

Table A.6: Measures and Summary Statistics for Process & Experiential Quality (2007-2013)

Code	Measure Name	Mean	SD	Obs.	2007	2008	2009	2010	2011	2012	2013
	Process Quality										
	Condition: AMI										
AMI-1	Patients Given Aspirin at Arrival	97.04	6.41	1654	○	○	○	○	○		
AMI-2	Patients Given Aspirin at Discharge	96.39	8.15	2169	○	○	○	○	○	○	○
AMI-3	Patients Given ACE Inhibitor or ARB for Left Ventricular Systolic Dysfunction (LVSD)	92.76	13.84	1460	○	○	○	○	○		
AMI-4	Patients Given Smoking Cessation Advice/Counseling	97.38	11.21	1393	○	○	○	○	○		
AMI-5	Patients Given Beta Blocker at Discharge	96.67	7.69	1630	○	○	○	○	○		
AMI-6	Patients Given Beta Blocker at Arrival	92.47	8.05	346	○						
AMI-7A	Patients Given Fibrinolytic Medication Within 30 Minutes Of Arrival	41.60	39.90	356	○	○	○	○	○	○	○
AMI-8A	Patients Given PCI Within 90 Minutes Of Arrival	87.28	15.01	1040	○	○	○	○	○	○	○
AMI-10	Patients Given a Prescription for a Statin at Discharge	96.15	6.70	820					○	○	○
OP-2	Fibrinolytic Therapy received within 30 minutes	54.08	33.86	206			○	○	○	○	○
OP-3b	Median Time to transfer patients for Acute Coronary Intervention	86.50	57.03	290			○	○	○	○	○
OP-4	Aspirin at Arrival	96.06	10.60	1067			○	○	○	○	○
	Condition: HF										
HF-1	Patients Given Discharge Instructions	87.08	16.82	2366	○	○	○	○	○	○	○
HF-2	Patients Given an Evaluation of Left Ventricular Systolic (LVS) Function	96.51	9.95	2377	○	○	○	○	○	○	○
HF-3	Patients Given ACE Inhibitor or ARB for Left Ventricular Systolic Dysfunction (LVSD)	93.40	9.23	2289	○	○	○	○	○	○	○
HF-4	Patients Given Smoking Cessation Advice/Counseling	96.61	9.42	1647	○	○	○	○	○		
	Condition: PN										
PN-1	Patients Given Oxygenation Assessment	99.51	1.68	705	○	○					
PN-2	Patients Assessed and Given Pneumococcal Vaccination	89.36	12.71	1717	○	○	○	○	○		
PN-3B	Patients Whose Initial Emergency Room Blood Culture Was Performed Prior To The Administration Of The First Hospital Dose Of Antibiotics	94.06	7.52	2360	○	○	○	○	○	○	○
PN-4	Patients Given Smoking Cessation Advice/Counseling	95.87	9.88	1689	○	○	○	○	○		
PN-5C	Patients Given Initial Antibiotic(s) within 6 Hours After Arrival	93.67	6.85	1694	○	○	○	○	○		
PN-6	Patients Given the Most Appropriate Initial Antibiotic(s)	92.59	8.31	2372	○	○	○	○	○	○	○
PN-7	Patients Assessed and Given Influenza Vaccination	87.29	14.29	1695	○	○	○	○	○		
	Experiential Quality										
Comp 1	Nurses communicated well	71.72	6.96	2265	○	○	○	○	○	○	○
Comp 2	Doctors communicated well	76.28	5.29	2265	○	○	○	○	○	○	○
Comp 3	Help received quickly	57.66	8.94	2265	○	○	○	○	○	○	○
Comp 4	Pain controlled well	65.54	6.32	2265	○	○	○	○	○	○	○
Comp 5	Staff explained medicines	56.54	6.71	2265	○	○	○	○	○	○	○
Comp 6	Given discharge instructions	80.11	5.60	2265	○	○	○	○	○	○	○

Notes: Some measures, which have too few ($n < 25$) cases for purposes of reliably predicting hospital performance, do not have scores, leading to a small number of observations (e.g., AMI-7A, OP-2).

Table A.7: Correlation Table for Lagged Dependent, Independent, and Control Variables

	Variable	n	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
(1)	Lag_ln(Total LOS)	1689	1														
(2)	Lag_ln(Total Cost)	1678	0.452	1													
(3)	Lag_ln(Cost of Ordering Test)	1672	0.024	0.639	1												
(4)	Lag_ln(Cost of Care Delivery)	1678	0.518	0.92	0.352	1											
(5)	WACVG	2004	0.426	0.205	-0.124	0.309	1										
(6)	WACVG_Underuse	2017	0.13	-0.093	-0.375	0.078	0.357	1									
(7)	WaCVG_Overuse	2017	0.164	-0.104	-0.413	0.086	0.4	0.889	1								
(8)	ProcQual	2356	0.053	0.078	0.193	-0.022	0.116	-0.002	0.004	1							
(9)	ExpeQual	2265	-0.346	-0.143	-0.026	-0.124	-0.263	-0.097	-0.124	-0.092	1						
(10)	Lag_ln(Bed Size)	2097	0.329	0.255	0.192	0.194	0.482	0.141	0.196	0.462	-0.391	1					
(11)	Lag_CMI	1972	0.113	0.193	0.164	0.16	0.317	0.131	0.126	0.433	-0.105	0.681	1				
(12)	Lag_Teacing Intensity	1972	0.304	0.222	-0.19	0.345	0.569	0.29	0.327	0.028	-0.299	0.385	0.289	1			
(13)	Lag_Wage Index	1954	0.45	0.519	-0.006	0.639	0.423	0.195	0.223	0.019	-0.298	0.264	0.12	0.569	1		
(14)	Lag_Outlier Adj Factor	1972	0.099	0.272	0.18	0.258	0.316	0.066	0.056	0.12	0.022	0.373	0.408	0.272	0.112	1	
(15)	Lag_OPDSH Adj Factor	1972	0.222	0.049	-0.208	0.126	0.353	0.185	0.19	0	-0.324	0.283	-0.026	0.612	0.445	0.113	1

Notes: $p < 0.05$ if $|r| > 0.048$. n is the number of hospital-level observations for six years (2008 to 2013).

Table A.8: Correlation Table for Alternative Practice Variation Measures

		Variable	n	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
(1)	WACVG(All,Hospital)	$WACVG^{H}_{all}$	2004	1																	
(2)	WACVD(All,Hospital)	$WACVD^{H}_{all}$	2004	0.837	1																
(3)	WACVG(All,County)	$WACVG^{C}_{all}$	2004	0.452	0.393	1															
(4)	WACVD(All,County)	$WACVD^{C}_{all}$	2004	0.439	0.402	0.85	1														
(5)	WACVG(All,StateCBSA)	$WACVG^{S}_{all}$	2004	0.336	0.233	0.595	0.532	1													
(6)	WACVD(All,StateCBSA)	$WACVD^{S}_{all}$	2004	0.368	0.269	0.591	0.678	0.844	1												
(7)	WACVG(Underuse,Hospital)	$WACVG^{H}_{under}$	2017	0.357	0.321	0.113	0.07	0.065	0.023	1											
(8)	WACVD(Underuse,Hospital)	$WACVD^{H}_{under}$	2030	0.368	0.35	0.132	0.087	0.08	0.039	0.985	1										
(9)	WACVG(Underuse,County)	$WACVG^{C}_{under}$	2017	0.205	0.219	-0.257	-0.231	-0.213	-0.195	0.257	0.258	1									
(10)	WACVD(Underuse,County)	$WACVD^{C}_{under}$	2030	0.209	0.228	-0.248	-0.228	-0.202	-0.188	0.261	0.19	0.994	1								
(11)	WACVG(Underuse,StateCBSA)	$WACVG^{S}_{under}$	2017	0.117	0.138	-0.204	-0.185	-0.133	-0.217	0.231	0.23	0.612	0.612	1							
(12)	WACVD(Underuse,StateCBSA)	$WACVD^{S}_{under}$	2030	0.097	0.123	-0.206	-0.187	-0.132	-0.218	0.219	0.177	0.598	0.609	0.986	1						
(13)	WACVG(Overuse,Hospital)	$WACVG^{H}_{over}$	2017	0.4	0.39	0.141	0.11	0.096	0.071	0.889	0.882	0.31	0.309	0.268	0.25	1					
(14)	WACVD(Overuse,Hospital)	$WACVD^{H}_{over}$	2030	0.41	0.433	0.176	0.139	0.119	0.095	0.851	0.874	0.308	0.259	0.254	0.205	0.959	1				
(15)	WACVG(Overuse,County)	$WACVG^{C}_{over}$	2017	0.096	0.061	0.548	0.473	0.344	0.334	0.112	0.111	-0.693	-0.692	-0.385	-0.383	0.093	0.092	1			
(16)	WACVD(Overuse,County)	$WACVD^{C}_{over}$	2030	0.116	0.093	0.558	0.488	0.355	0.347	0.111	0.137	-0.684	-0.682	-0.381	-0.387	0.096	0.129	0.993	1		
(17)	WACVG(Overuse,StateCBSA)	$WACVG^{S}_{over}$	2017	0.053	-0.018	0.405	0.356	0.538	0.533	-0.03	-0.036	-0.533	-0.535	-0.754	-0.747	-0.044	-0.046	0.547	0.541	1	
(18)	WACVD(Overuse,StateCBSA)	$WACVD^{S}_{over}$	2030	0.074	0.009	0.429	0.38	0.549	0.548	-0.028	-0.01	-0.524	-0.526	-0.736	-0.723	-0.04	-0.014	0.549	0.555	0.988	1

Notes: $p < 0.05$ if $|r| > 0.061$. n is the number of hospital-level observations for six years (2008 to 2013).

A.8 Charge Components by Condition

Table A.9 lists the percentage of non-zero observations (i.e., the percentage of patients who received at least a single service that incurs a non-zero amount of charges), and the percentage of total charges for each state, for each charge component by condition.

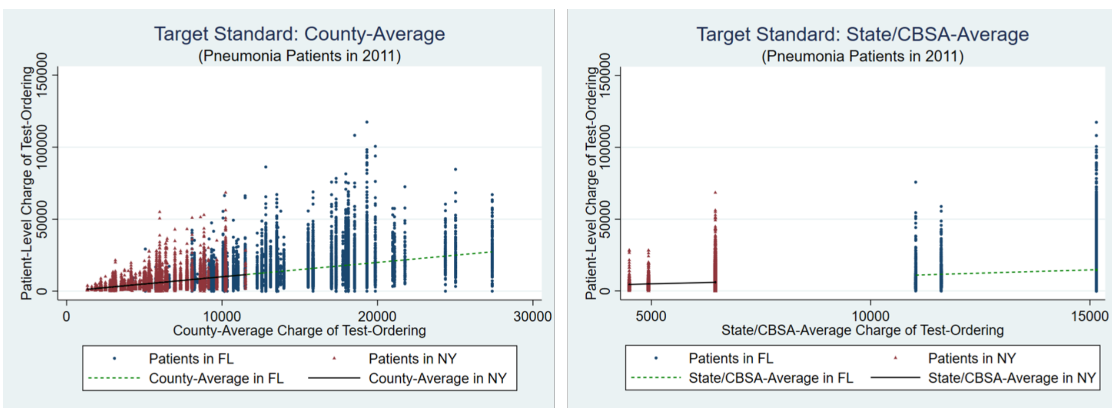
Table A.9: Charge Components by Condition

Condition	AMI			
Charge Name	% Non-zero Obs.	% Totchg FL	% Totchg NY	% Totchg All
Emergency Room Charge	85.00	6.32	5.74	6.05
Room Charges	54.01	4.17	28.52	15.64
Laboratory Charges	98.54	21.62	13.04	17.58
Radiology and Other Imaging	94.28	9.31	5.92	7.72
Intensive Care Charges	45.88	10.94	9.96	10.48
Coronary Care Charges	25.24	4.57	8.81	6.56
Pharmacy Charges	94.92	12.55	5.85	9.39
Medical and Surgical Supply	75.69	4.14	1.89	3.08
Respiratory Services	48.54	2.98	2.10	2.56
Cardiology Charges	77.73	16.41	10.38	13.57
Condition	HF			
Charge Name	% Non-zero Obs.	% Totchg FL	% Totchg NY	% Totchg All
Emergency Room Charge	87.91	7.54	6.33	6.99
Room Charges	64.94	7.74	43.86	24.01
Laboratory Charges	98.63	25.1	13.74	19.95
Radiology and Other Imaging	97.56	10.14	5.43	8.01
Intensive Care Charges	32.16	11.24	4.29	8.10
Coronary Care Charges	17.22	4.97	5.10	5.03
Pharmacy Charges	95.39	12.24	4.78	8.86
Medical and Surgical Supply	66.69	3.4	1.39	2.50
Respiratory Services	60.54	4.43	2.81	3.70
Cardiology Charges	55.39	5.37	4.33	4.9
Condition	PN			
Charge Name	% Non-zero Obs.	% Totchg FL	% Totchg NY	% Totchg All
Emergency Room Charge	87.77	7.10	6.48	6.81
Room Charges	83.04	11.72	47.15	28.56
Laboratory Charges	98.52	20.53	13.12	17.01
Radiology and Other Imaging	96.95	13.54	7.46	10.65
Intensive Care Charges	19.67	7.32	2.08	4.83
Coronary Care Charges	7.51	2.94	1.48	2.24
Pharmacy Charges	95.76	18.33	7.9	13.37
Medical and Surgical Supply	66.52	3.58	1.76	2.71
Respiratory Services	73.23	6.32	4.33	5.37
Cardiology Charges	21.79	2.11	1.69	1.91

A.9 Robustness Check: Alternative Measure of Practice Variation

In calculating the underuse/overuse variation of test-ordering practice in the manuscript, we use "hospital-average" test-ordering charge as a presumed target standard. "County-average" and "State/CBSA[1]-average" could also be worthwhile standards, as illustrated for pneumonia patients in 2011, as shown in the example in Figure A.2. Indeed, a recent mandatory bundled payment program operated by CMS, namely Comprehensive Care for Joint Replacement (CJP) Model, adopts a similar approach. In the CJP program, the target price for a care-episode bundle is set prospectively and reflects a blend of hospital-specific and regional data,[2] suggesting the possibility of using a target standard obtained from a broader regional boundary than a single hospital.

Figure A.2: Variation in Test-Ordering Practice, Target: County-Average (Left) and State/CBSA-Average (Right)



We provide results of using the alternative practice variation metrics with different targets. As listed in Table A.10 for Hypothesis 1 and Tables A.11 and A.12 for Hypothesis 2, we observe largely consistent results. Practice variation measures using the State/CBSA-average as a target standard tend to show relatively weak evidence (either smaller magnitudes in coefficients or insignificant coefficients), possibly implying that using an unnecessarily large regional bound-

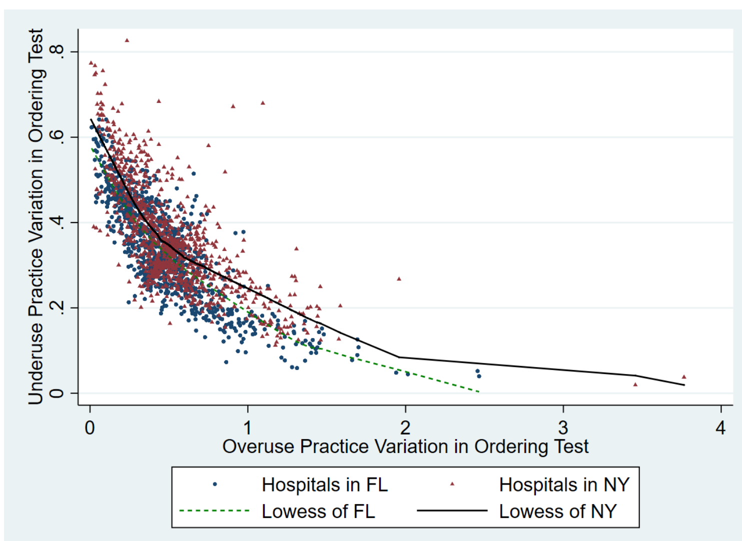
[1]Core-Based Statistical Area refers collectively to both metropolitan statistical areas and micropolitan areas.

[2]http://www.singletrackanalytics.com/blog/15-11-23/top-ten-things-you-need-know-now-medicare-cjr-program-final

ary may diminish the association between the practice variation and operational performance that could be identified with a target standard having a proper regional boundary.

Lastly, we observe an interesting pattern between the underuse- and overuse-variation in test-ordering practice when county-average is used as a target standard (see Figure A.3). They are inversely proportional to each other, meaning that a hospital that tends to overuse laboratory/radiology tests is less likely to underuse the test, which makes sense. By performing an efficient frontier analysis, one may identify Pareto-optimal hospitals for test-ordering practice. Examining drivers of their detailed test-ordering practices could facilitate practitioners and policy-makers to identify "best practices," suggesting an opportunity for worthwhile future research.

Figure A.3: Underuse- vs. Overuse-Variation in Test-Ordering Practice



A.10 Robustness Check: Results by Condition

Figure A.4 shows histograms of the practice variation measure constructed by condition, i.e., $\text{WACVG}^{H}_{\text{all},j}$ where $j \in \{AMI, HF, PN\}$. We observe similar distributions of practice variation across the hospitals for each condition. We examine our hypotheses condition-by-condition and consistently find supporting evidence as listed in Table A.13 and Table A.14.

Table A.10: Results of System GMM (Alternative Practice Variation Measures for Testing H1)

Dep Var	Length of Stay							
WACV Type	$\text{EPVRG}^{H}_{\text{all}}$	$\text{WACVG}^{H}_{\text{all}}$	$\text{WACVG}^{C}_{\text{all}}$	$\text{WACVG}^{S}_{\text{all}}$	$\text{EPVRD}^{H}_{\text{all}}$	$\text{WACVD}^{H}_{\text{all}}$	$\text{WACVD}^{C}_{\text{all}}$	$\text{WACVD}^{S}_{\text{all}}$
	(M1)	(M2)	(M3)	(M4)	(M5)	(M6)	(M7)	(M8)
ProcQuality (PQ)	0.001	−0.073	0.007	0.048	0.032	0.043	0.028	0.010
	(0.033)	(0.047)	(0.057)	(0.037)	(0.035)	(0.034)	(0.020)	(0.011)
ExpQuality (EQ)	−0.198*	−0.549***	−0.683**	−0.220	−0.433**	−0.546***	−0.219*	−0.143***
	(0.103)	(0.205)	(0.334)	(0.203)	(0.178)	(0.188)	(0.134)	(0.055)
WACV	0.312***	0.635***	0.591***	0.261***	0.282***	0.456***	0.148***	0.055***
	(0.034)	(0.070)	(0.117)	(0.076)	(0.037)	(0.062)	(0.033)	(0.013)
PQ*WACV	0.003	0.130	−0.031	−0.068	−0.034	−0.073	−0.045*	−0.016
	(0.033)	(0.084)	(0.100)	(0.050)	(0.035)	(0.058)	(0.023)	(0.010)
EQ*WACV	0.246**	0.920**	1.285**	0.227	0.384**	0.927***	0.252*	0.100**
	(0.112)	(0.369)	(0.564)	(0.290)	(0.185)	(0.313)	(0.144)	(0.042)
Dep $\text{Var}_{(t-1)}$	0.536***	0.551***	0.386***	0.692***	0.586***	0.584***	0.623***	0.768***
	(0.055)	(0.064)	(0.119)	(0.093)	(0.051)	(0.050)	(0.081)	(0.068)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year Dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1527	1527	1527	1527	1527	1527	1527	1527
Hospitals	324	324	324	324	324	324	324	324
Instruments	65	65	65	65	65	65	65	65
AR(1) (p-value)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
AR(2) (p-value)	(0.261)	(0.205)	(0.417)	(0.093)	(0.122)	(0.160)	(0.068)	(0.060)
Hansen test (p-value)	(0.182)	(0.238)	(0.077)	(0.091)	(0.108)	(0.184)	(0.418)	(0.177)
Diff.-in-Hansen test (p-value)	(0.326)	(0.286)	(0.755)	(0.410)	(0.524)	(0.475)	(0.606)	(0.730)

Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: The results are based on a system GMM model (Arellano and Bond, 1991; Blundell and Bond, 1998) estimated as in Equation (2.11). Standard errors are corrected for heteroskedasticity. AR(1) and AR(2) are tests for first-order and second-order serial correlation in the first-differenced residuals, under the null of no serial correlation. The Hansen test of overidentification is under the null that all instruments are valid. The difference-in-Hansen test of exogeneity is under the null that instruments used for the equations in levels are exogenous.

Table A.11: Results of System GMM (Alternative WACVG Practice Variation Measures for Testing H2)

Dep Var	Total Cost			Cost of Ordering Test			Cost of Delivering Care		
WACV Target	Hospital	County	State	Hospital	County	State	Hospital	County	State
	(M1)	(M2)	(M3)	(M4)	(M5)	(M6)	(M7)	(M8)	(M9)
ProcQuality (PQ)	−0.004	−0.005	−0.014	−0.029	−0.013	0.006	−0.035	−0.002	−0.013**
	(0.017)	(0.017)	(0.017)	(0.023)	(0.021)	(0.006)	(0.056)	(0.011)	(0.006)
ExpQuality (EQ)	−0.168*	−0.100	0.023	−0.221	0.039	0.069	−0.476***	−0.194**	−0.069
	(0.091)	(0.132)	(0.084)	(0.174)	(0.118)	(0.057)	(0.227)	(0.097)	(0.058)
WACVG_Under	0.560*	0.248**	0.190*	−0.288**	−0.433***	−0.142**	0.640***	0.466***	0.140***
	(0.290)	(0.120)	(0.099)	(0.144)	(0.147)	(0.065)	(0.212)	(0.070)	(0.045)
WACVG_Over	0.091	−0.042	−0.060	0.258**	0.066**	0.056***	−0.203	−0.037	−0.040*
	(0.077)	(0.061)	(0.051)	(0.126)	(0.032)	(0.020)	(0.206)	(0.027)	(0.024)
PQ*WACVG_Under	−0.230	0.007	0.037	0.111	0.268*	0.012	0.107	−0.038	0.017
	(0.166)	(0.052)	(0.041)	(0.074)	(0.158)	(0.024)	(0.177)	(0.032)	(0.013)
PQ*WACVG_Over	0.009	−0.051	−0.044	−0.114	0.027	−0.004	0.275	0.006	0.013
	(0.034)	(0.053)	(0.035)	(0.095)	(0.031)	(0.006)	(0.181)	(0.037)	(0.009)
EQ*WACVG_Under	0.299	0.294	−0.033	0.727	0.983	−0.065	1.588**	0.469*	0.193*
	(0.671)	(0.378)	(0.201)	(0.527)	(1.183)	(0.246)	(0.729)	(0.273)	(0.108)
EQ*WACVG_Over	0.353**	0.005	0.161	0.165	−0.141	−0.086	1.433***	0.212**	−0.059
	(0.174)	(0.123)	(0.169)	(0.270)	(0.141)	(0.059)	(0.553)	(0.103)	(0.111)
Dep Var$_{(t-1)}$	0.758***	0.510***	0.612***	0.931***	0.860***	0.782***	0.735***	0.666***	0.729***
	(0.068)	(0.170)	(0.153)	(0.032)	(0.085)	(0.060)	(0.072)	(0.148)	(0.071)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year Dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1519	1519	1519	1519	1519	1519	1519	1519	1519
Hospitals	324	324	324	324	324	324	324	324	324
Instruments	83	83	83	83	83	83	83	83	83
AR(1) (p-value)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
AR(2) (p-value)	(0.240)	(0.475)	(0.041)	(0.571)	(0.240)	(0.368)	(0.977)	(0.481)	(0.101)
Hansen test (p-value)	(0.201)	(0.078)	(0.507)	(0.167)	(0.066)	(0.054)	(0.083)	(0.206)	(0.255)
Diff.-in-Hansen test (p-value)	(0.376)	(0.593)	(0.464)	(0.387)	(0.277)	(0.180)	(0.340)	(0.771)	(0.386)

Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: The results are based on a system GMM model (Arellano and Bond, 1991; Blundell and Bond, 1998) estimated as in Equation (2.11). Standard errors are corrected for heteroskedasticity. AR(1) and AR(2) are tests for first-order and second-order serial correlation in the first-differenced residuals, under the null of no serial correlation. The Hansen test of overidentification is under the null that all instruments are valid. The difference-in-Hansen test of exogeneity is under the null that instruments used for the equations in levels are exogenous.

Table A.12: Results of System GMM (Alternative WACVD Practice Variation Measures for Testing H2)

Dep Var	Total Cost			Cost of Ordering Test			Cost of Delivering Care		
WACV Target	Hospital	County	State	Hospital	County	State	Hospital	County	State
	(M1)	(M2)	(M3)	(M4)	(M5)	(M6)	(M7)	(M8)	(M9)
ProcQuality (PQ)	−0.007	−0.004	−0.018	−0.035	0.011	0.008	−0.025	−0.010	−0.039**
	(0.014)	(0.005)	(0.017)	(0.027)	(0.007)	(0.006)	(0.026)	(0.006)	(0.017)
ExpQuality (EQ)	−0.108*	−0.011	0.037	−0.251	0.037	−0.005	−0.580***	−0.138**	−0.068
	(0.056)	(0.052)	(0.082)	(0.162)	(0.062)	(0.053)	(0.151)	(0.062)	(0.104)
WACVD_Under	0.382*	0.353***	−0.105	−0.280*	−0.288***	−0.137**	0.313**	0.435***	0.262***
	(0.200)	(0.037)	(0.064)	(0.165)	(0.070)	(0.055)	(0.140)	(0.039)	(0.099)
WACVD_Over	0.034	−0.005	0.050	0.407**	0.033**	0.056***	−0.044	−0.014	−0.065
	(0.047)	(0.016)	(0.044)	(0.169)	(0.016)	(0.018)	(0.059)	(0.016)	(0.042)
PQ*WACVD_Under	−0.049	0.013	0.031	0.118	0.139***	0.027	−0.217	0.017	0.080**
	(0.132)	(0.015)	(0.038)	(0.081)	(0.049)	(0.017)	(0.136)	(0.015)	(0.031)
PQ*WACVD_Over	0.020	−0.002	0.003	−0.136*	−0.006	−0.008	0.188***	0.004	0.027
	(0.026)	(0.008)	(0.018)	(0.081)	(0.010)	(0.006)	(0.058)	(0.010)	(0.021)
EQ*WACVD_Under	0.715	0.153	0.050	0.770	0.527	0.040	1.395*	0.601***	0.126
	(0.554)	(0.135)	(0.193)	(0.497)	(0.397)	(0.190)	(0.735)	(0.157)	(0.182)
EQ*WACVD_Over	0.231**	−0.073	−0.076	−0.538	−0.095	0.002	0.308	0.104**	−0.167
	(0.107)	(0.049)	(0.134)	(0.488)	(0.070)	(0.052)	(0.288)	(0.050)	(0.131)
Dep Var$_{(t-1)}$	0.756***	0.442***	0.719***	0.963***	0.868***	0.797***	0.680***	0.424***	0.643***
	(0.055)	(0.045)	(0.083)	(0.041)	(0.056)	(0.061)	(0.041)	(0.041)	(0.151)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year Dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1519	1519	1519	1519	1519	1519	1519	1519	1519
Hospitals	324	324	324	324	324	324	324	324	324
Instruments	83	83	83	83	83	83	83	83	83
AR(1) (p-value)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
AR(2) (p-value)	(0.360)	(0.364)	(0.399)	(0.280)	(0.260)	(0.324)	(0.698)	(0.278)	(0.104)
Hansen test (p-value)	(0.201)	(0.155)	(0.170)	(0.135)	(0.061)	(0.304)	(0.052)	(0.185)	(0.234)
Diff.-in-Hansen test (p-value)	(0.321)	(0.259)	(0.305)	(0.190)	(0.204)	(0.265)	(0.155)	(0.739)	(0.084)

Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: The results are based on a system GMM model (Arellano and Bond, 1991; Blundell and Bond, 1998) estimated as in Equation (2.11). Standard errors are corrected for heteroskedasticity. AR(1) and AR(2) are tests for first-order and second-order serial correlation in the first-differenced residuals, under the null of no serial correlation. The Hansen test of overidentification is under the null that all instruments are valid. The difference-in-Hansen test of exogeneity is under the null that instruments used for the equations in levels are exogenous.

Figure A.4: Practice Variation by Condition

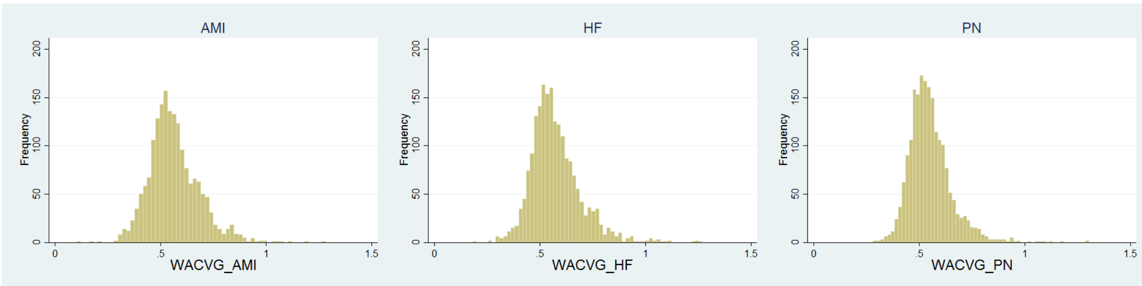


Table A.13: Results of System GMM (Testing H1 by Hospital-Condition)

Dep Var$_j$	Length of Stay		
Medical Condition	AMI	HF	PN
	(M1)	(M2)	(M3)
ProcQuality$_j$ (PQ$_j$)	-0.043	-0.047^{*}	-0.030
	(0.044)	(0.028)	(0.058)
ExpQuality (EQ)	-0.959^{**}	-0.531^{**}	-0.590^{**}
	(0.382)	(0.224)	(0.229)
WACVG$^{H}_{\text{all},j}$	0.340^{***}	0.451^{***}	0.521^{***}
	(0.081)	(0.054)	(0.057)
PQ$_j$*WACVG$^{H}_{\text{all},j}$	0.095	0.062	0.046
	(0.078)	(0.053)	(0.103)
EQ*WACVG$^{H}_{\text{all},j}$	1.245^{*}	0.770^{**}	1.033^{**}
	(0.692)	(0.391)	(0.422)
Dep Var$_{(t-1)}$	0.869^{***}	0.575^{***}	0.428^{***}
	(0.119)	(0.060)	(0.064)
Controls	Yes	Yes	Yes
Year Dummies	Yes	Yes	Yes
Observations	1390	1344	1482
Hospitals	313	321	322
Instruments	65	65	65
AR(1) (p-value)	(0.000)	(0.000)	(0.000)
AR(2) (p-value)	(0.810)	(0.694)	(0.091)
Hansen test (p-value)	(0.182)	(0.130)	(0.433)
Diff.-in-Hansen test (p-value)	(0.104)	(0.101)	(0.415)

Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.
Subscript j represents that condition-specific measures are used.

Table A.14: Results of System GMM (Testing H2 by Hospital-Condition)

Dep Var$_j$	Cost of Care-Delivery		
Medical Condition	AMI	HF	PN
	(M1)	(M2)	(M3)
ProcQuality$_j$ (PQ$_j$)	0.056	0.005	0.051
	(0.036)	(0.014)	(0.072)
ExpQuality (EQ)	-0.608^{**}	-0.143^{**}	-0.691^{**}
	(0.243)	(0.056)	(0.342)
WACVG$^H_{\text{under},j}$	0.750^{***}	0.322^{***}	0.731^{**}
	(0.191)	(0.087)	(0.362)
WACVG$^H_{\text{over},j}$	0.086	-0.022	-0.207
	(0.154)	(0.064)	(0.319)
PQ$_j$*WACVG$^H_{\text{unver},j}$	-0.172	-0.046	-0.156
	(0.106)	(0.049)	(0.221)
PQ$_j$*WACVG$^H_{\text{over},j}$	-0.138	0.019	0.192
	(0.095)	(0.044)	(0.247)
EQ*WACVG$^H_{\text{under},j}$	1.885^{**}	0.370^{**}	2.224^{**}
	(0.745)	(0.186)	(1.038)
EQ*WACVG$^H_{\text{over},j}$	0.753	-0.008	-0.173
	(0.654)	(0.097)	(0.686)
Dep Var$_{(t-1),j}$	0.743^{***}	0.753^{***}	0.620^{***}
	(0.082)	(0.029)	(0.104)
Controls	Yes	Yes	Yes
Year Dummies	Yes	Yes	Yes
Observations	1390	1344	1482
Hospitals	313	321	322
Instruments	83	83	83
AR(1) (p-value)	(0.000)	(0.000)	(0.000)
AR(2) (p-value)	(0.983)	(0.661)	(0.062)
Hansen test (p-value)	(0.309)	(0.360)	(0.160)
Diff.-in-Hansen test (p-value)	(0.146)	(0.376)	(0.216)

Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.
Subscript j represents that condition-specific measures are used.

A.11 Robustness Check: Alternative Measure of Process Quality

Anecdotally, dealing with process quality (PQ) measures imposes a considerable burden on physician practices in terms of understanding the measures, collecting and reporting performance data, and understanding performance reports from third party payers, but the extent of that burden has not been quantified (Casalino et al., 2016). Although there is much to gain from quality measurement, the current PQ measurement system is far from efficient and contributes to negative physician attitudes toward quality measures (Meltzer and Chung, 2014). To capture the burden of measuring and reporting process quality, a measure that is being updated every year as listed in Table A.6, we measure process quality burden (PQB) as below to address any association between the burden of quality measurement and its impact on operational performance:

$$PQB^{o}_{it} = \frac{\sum_{j \in J} \sum_{m \in M_j} |\Delta N_{ijmt}| \cdot \Delta q_{ijmt}}{\sum_{j \in J} \sum_{m \in M_j} |\Delta N_{ijmt}|}, \tag{A.1}$$

where $\Delta N_{ijmt} = N_{ijmt} - N_{ijm,t-1}$ and $\Delta q_{ijmt} = q_{ijmt} - q_{ijm,t-1}$.

Note that if a measure m is newly added in year t, then $q_{ijmt'} = 0$ for $t' < t$. Similarly, if the measure m is topped-out in year t, then $q_{ijmt''} = 0$ for $t'' \geq t$. In this way, we capture the increased or decreased burden of measuring and reporting a specific process quality metric. Instead of using an indicator variable approach, which cannot capture subtle variations over time, we adopt this approach to reflect heterogeneous reactions of hospitals to PQ measure updates. For example, one hospital might have a much larger relevant patient sample, compared to another hospital, in applying a new PQ metric, implying different levels of burden.

We find consistent results even after replacing the original PQ measure with the PQB measure, as listed in Table A.15. PQB (as well as PQ) is not significantly associated with performance either directly or as a moderator of practice variation.

Table A.15: Results of System GMM (Testing H1 with an Alternative Process Quality Capturing Burden of Measurement)

Dep Var	Length of Stay							
WACV Type	$\text{EPVRG}^H_{\text{all}}$	$\text{WACVG}^H_{\text{all}}$	$\text{WACVG}^C_{\text{all}}$	$\text{WACVG}^S_{\text{all}}$	$\text{EPVRD}^H_{\text{all}}$	$\text{WACVD}^H_{\text{all}}$	$\text{WACVD}^C_{\text{all}}$	$\text{WACVD}^S_{\text{all}}$
	(M1)	(M2)	(M3)	(M4)	(M5)	(M6)	(M7)	(M8)
PQ_Burden (PQB)	0.080	0.004	−0.032	−0.088	0.232*	0.230	−0.010	−0.083
	(0.130)	(0.121)	(0.116)	(0.060)	(0.138)	(0.154)	(0.086)	(0.082)
ExpQuality (EQ)	−0.141**	−0.183**	−0.615**	−0.022	−0.326***	−0.475***	−0.172**	−0.159**
	(0.070)	(0.086)	(0.283)	(0.145)	(0.110)	(0.177)	(0.083)	(0.063)
WACV	0.275***	0.462***	0.559***	0.206***	0.269***	0.470***	0.102***	0.041***
	(0.028)	(0.047)	(0.110)	(0.059)	(0.028)	(0.050)	(0.019)	(0.011)
PQB*WACV	−0.049	−0.013	0.039	0.134**	−0.200	−0.310	0.012	0.075
	(0.114)	(0.191)	(0.149)	(0.068)	(0.132)	(0.244)	(0.069)	(0.051)
EQ*WACV	0.160*	0.255*	1.204**	−0.081	0.233**	0.700**	0.168*	0.092*
	(0.093)	(0.154)	(0.482)	(0.226)	(0.109)	(0.305)	(0.087)	(0.055)
Dep $\text{Var}_{(t-1)}$	0.566***	0.631***	0.400***	0.729***	0.608***	0.577***	0.696***	0.700***
	(0.057)	(0.042)	(0.110)	(0.073)	(0.044)	(0.053)	(0.063)	(0.082)
Year Dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1527	1527	1527	1527	1527	1527	1527	1527
Hospitals	324	324	324	324	324	324	324	324
Instruments	59	59	59	59	59	59	59	59
AR(1) (p-value)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
AR(2) (p-value)	(0.099)	(0.071)	(0.410)	(0.102)	(0.053)	(0.073)	(0.043)	(0.063)
Hansen test (p-value)	(0.332)	(0.082)	(0.083)	(0.053)	(0.206)	(0.157)	(0.056)	(0.153)
Diff.-in-Hansen test (p-value)	(0.382)	(0.247)	(0.407)	(0.064)	(0.304)	(0.409)	(0.209)	(0.147)

Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

A.12 Post-hoc Analysis: Patient Outcomes as Dependent Variables

While our main dependent variables of interest are operational performance measures such as patient length-of-stay and care-delivery cost per capita, whether the practice variation relates to patient outcomes is also an important question, considering "The Triple Aim: Care, Health, and Cost" developed by the Institute of Healthcare Improvement.

We hence test the impact of practice variation on a hospital's readmission rate and mortality rate. While lower practice variation is associated with lower mortality rate as M6 in Table A.16 implies, we could not find any significant relationship between the practice variation and readmission rate as in M3. Note, however, that our readmission and mortality rates impose some limitations. Our HCUP data can track patient revisits only within a hospital (i.e., we are not able to track patient readmission across hospitals) and thus our readmission rate is not complete.[3] Similarly, our mortality rate is also not complete since we can only track patient deaths that occurred during a hospital stay (unlike the CMS measure that tracks patient deaths even after discharge up to 30 days). We thus leave the delicate examination of this relationship as future work.

A.13 Post-hoc Analysis: Determinants of Practice Variation

A main focus of bundled payment contracts is to standardize care processes by changing the ways physicians and hospitals deliver care. In this section, we report post-hoc analyses that investigate another important question, concerning what factors lead to different levels of practice variation across hospitals. In particular, we investigate (1) whether the time-varying hospital-level characteristics are determinants of practice variation, and (2) whether the practice variation originates from physician-level personal behavior or organizational-level environments. To address the first issue, we estimate a dynamic panel model with the practice variation measure as a dependent variable. The second issue is addressed in a descriptive sense using a t-test as follows.

[3] Our weighted average readmission rate across AMI, HF, and PN is around 11.5% on average (SD: 0.022, Max: 0.152) while the CMS hospital compare data shows around 21% on average (SD: 0.02, Max: 0.29) (Senot et al., 2015).

Table A.16: Hospital-Level Results of System GMM (WACVG, Dep: ReadmRate, MortRate)

Dep Var	Readmission Rate			Mortality Rate		
	(M1)	(M2)	(M3)	(M4)	(M5)	(M6)
ProcQuality (PQ)	0.001	0.001	0.006	−0.000	0.000	-0.020^{**}
	(0.001)	(0.001)	(0.011)	(0.001)	(0.001)	(0.009)
ExpQuality (EQ)	0.004	−0.004	−0.055	0.004	0.005	−0.019
	(0.007)	(0.005)	(0.064)	(0.009)	(0.010)	(0.055)
WACVG$^{H}_{\textbf{all}}$		0.026^{***}	0.045		0.026^{**}	0.031^{***}
		(0.007)	(0.048)		(0.011)	(0.011)
PQ*WACVG$^{H}_{\textbf{all}}$			−0.005			0.037^{**}
			(0.019)			(0.017)
EQ*WACVG$^{H}_{\textbf{all}}$			0.118			0.033
			(0.112)			(0.093)
Teaching Intensity	0.003	−0.002	0.004	0.011^{***}	0.005	−0.000
	(0.004)	(0.003)	(0.011)	(0.004)	(0.004)	(0.006)
Bed Size	−0.001	-0.002^{**}	-0.003^{*}	−0.000	−0.001	−0.001
	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)
Case Mix Index	-0.014^{***}	-0.014^{***}	-0.023^{***}	0.001	0.000	0.001
	(0.003)	(0.003)	(0.005)	(0.003)	(0.003)	(0.004)
Wage Index	−0.005	-0.008^{**}	−0.005	0.006^{*}	0.004	0.004
	(0.004)	(0.003)	(0.005)	(0.003)	(0.003)	(0.004)
OPDSH Adj Factor	0.008	0.006	0.015^{**}	−0.006	−0.005	−0.003
	(0.006)	(0.005)	(0.007)	(0.004)	(0.004)	(0.005)
Outlier Adj Factor	−0.013	−0.012	−0.020	-0.019^{*}	-0.026^{**}	-0.028^{**}
	(0.010)	(0.009)	(0.014)	(0.011)	(0.011)	(0.013)
Dep Var$_{(t-1)}$	0.044	0.017	−0.011	0.539^{***}	0.604^{***}	0.601^{***}
	(0.035)	(0.024)	(0.119)	(0.090)	(0.091)	(0.104)
Year Dummies	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1527	1527	1527	1527	1527	1527
Hospitals	324	324	324	324	324	324
Instruments	41	49	65	41	49	65
AR(1) (p-value)	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)
AR(2) (p-value)	(0.471)	(0.218)	(0.351)	(0.047)	(0.069)	(0.104)
Hansen test (p-value)	(0.060)	(0.119)	(0.148)	(0.062)	(0.105)	(0.347)
Diff.-in-Hansen test (p-value)	(0.050)	(0.091)	(0.062)	(0.320)	(0.131)	(0.321)

Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: The results are based on a system GMM model (Arellano and Bond, 1991; Blundell and Bond, 1998) estimated as in Equation (2.11). Standard errors are corrected for heteroskedasticity. AR(1) and AR(2) are tests for first-order and second-order serial correlation in the first-differenced residuals, under the null of no serial correlation. The Hansen test of overidentification is under the null that all instruments are valid. The difference-in-Hansen test of exogeneity is under the null that instruments used for the equations in levels are exogenous.

A.13.1 Determinants of Practice Variation: Hospital Characteristics

Our main analysis focused on examining the impact of practice variation on hospital operational performance. The underlying assumption is that the time-varying hospital characteristics that are used as proxies for the hospital's operating environment (i.e., teaching intensity, bed size, case mix index, wage index, operating disproportionate share hospital payment adjustment factor, and outlier adjustment factor) are the determinants of practice variation. In other words, the exogenous components of those characteristics are assumed to have a causal effect on practice variation (Wintoki et al., 2012). Previous studies provide empirical evidence showing that this is the case (e.g., Lake and Friese, 2006), but not all of these studies control for potential endogeneity that current hospital characteristics can be related to past levels of practice variation. Thus, we examine whether hospital characteristics are actual determinants of practice variation by applying the dynamic panel model (i.e., system GMM estimator) and further by comparing the results with the static model (i.e., OLS estimator). An empirical model we estimate is below:

$$x_{it} = \alpha + \sum_{p} \lambda_p \cdot x_{it-p} + \gamma \cdot \mathbf{Z_{it}} + \eta_i + \epsilon_{it} \tag{A.2}$$

where x is practice variation and Z_{it} is a vector of hospital characteristics including operational performance. Note that unlike our previous main model, we regress hospital characteristics on practice variation, while controlling for lagged practice variation. Table A.17 shows the results. Models (1) and (2) are the results of static[4] OLS and system GMM, respectively, with contemporaneous process quality and experiential quality, while Model (3) is with lagged quality measures. Models (4) to (6) are similar to Models (1) to (3) except that the WACVD version of practice variation is used, instead of WACVG. The results show that neither contemporaneous nor lagged quality metrics relate to the magnitude of practice variation. Instead, hospitals with higher teaching intensity, larger bed size, located in areas with a higher wage index, and outlier adjustment factors tend to have higher levels of practice variation. We discuss the implications in the next section, while here we focus more on comparing the results between static OLS and dynamic panel GMM

[4]Static in a sense that lagged practice variable is not included in the model.

estimates.

According to Model (2), the system GMM results are similar to the static OLS estimates (i.e., Model (1)) even after controlling for simultaneity, time-invariant unobserved heterogeneity, and the possible impact of past practice variation on current hospital characteristics. Thus, if we are interested in the effect of hospital characteristics on practice variation (a "structure-performance" relationship), then the relationship between present values of the explanatory variables and past values of the dependent variables may be less important than the relationship in our main model (i.e., the effect of practice variation on operational performance, or "performance-performance" relationship). Indeed, the explanatory variables (e.g., teaching intensity and bed size) are not determined by past dependent variables (i.e., practice variation). If any link exists from past practice variation to current hospital characteristics, then that will be indirect via the effect of practice variation on operational performance. While a strong relationship is reasonable between past hospital characteristics and current levels of practice variation, the reverse argument is weaker. Hospital managers usually do not increase current teaching intensity nor bed size because they have higher past levels of practice variation. Thus, as Table A.17 suggests, we have similar results from either OLS or system GMM estimates when we examine the effect of hospital characteristics on practice variation.

Meanwhile, it is also notable to examine the importance of controlling for both time-invariant unobservable heterogeneity and the dynamic relation between current practice variation and past hospital operational performance. For example, the estimated magnitudes of the effect of teaching intensity, bed size, wage index, and the outlier adjustment factor on practice variation from system GMM are smaller than those from static OLS estimates. This finding indicates possible upward biases from static OLS estimates due to the combination of unobservable heterogeneity and the endogeneity coming from the impact of past practice variation on current hospital characteristics. In Appendix A.13.3, we also investigate whether the level of practice variation varies across the hospital controls, which are static in most cases (i.e., governmental, non-profit, and for-profit).

Table A.17: Quality Measures as Antecedent of Practice Variation Measures (Dep: WACVG, WACVD)

Dep Var	WACVG_Hos			WACVD_Hos		
Estimation	OLS	System GMM		OLS	System GMM	
	(M1)	(M2)	(M3)	(M4)	(M5)	(M6)
ProcQuality (PQ)	-0.002	-0.002		-0.002	-0.000	
	(0.002)	(0.004)		(0.002)	(0.004)	
ExpQuality (EQ)	-0.001	-0.009		0.007	0.018	
	(0.011)	(0.041)		(0.012)	(0.040)	
PQ$_{(t-1)}$			0.002			0.002
			(0.005)			(0.005)
EQ$_{(t-1)}$			0.023			0.022
			(0.066)			(0.070)
Teaching Intensity	0.143^{***}	0.101^{***}	0.099^{***}	0.153^{***}	0.057^{**}	0.043^{*}
	(0.021)	(0.034)	(0.038)	(0.024)	(0.025)	(0.025)
Bed Size	0.024^{***}	0.019^{***}	0.018^{**}	0.012^{*}	0.010^{**}	0.008
	(0.006)	(0.006)	(0.007)	(0.007)	(0.005)	(0.006)
Case Mix Index	0.019	0.009	0.001	0.026	0.001	-0.004
	(0.016)	(0.015)	(0.015)	(0.020)	(0.016)	(0.018)
Wage Index	0.065^{***}	0.073^{***}	0.077^{***}	0.055^{**}	0.068^{***}	0.067^{***}
	(0.022)	(0.019)	(0.023)	(0.025)	(0.018)	(0.021)
OPDSH Adj Factor	-0.030	-0.021	-0.012	-0.003	0.005	0.004
	(0.022)	(0.022)	(0.024)	(0.028)	(0.020)	(0.021)
Outlier Adj Factor	0.148^{***}	0.141^{***}	0.120^{**}	0.092^{*}	0.074	0.065
	(0.045)	(0.046)	(0.059)	(0.053)	(0.061)	(0.076)
Total LOS$_{(t-1)}$	0.065^{***}	-0.007	-0.013	0.081^{***}	-0.043^{*}	-0.056^{**}
	(0.016)	(0.033)	(0.037)	(0.021)	(0.024)	(0.023)
Total Cost$_{(t-1)}$	-0.016^{*}	-0.014^{*}	-0.014^{*}	-0.023^{*}	-0.015^{**}	-0.013^{**}
	(0.009)	(0.008)	(0.009)	(0.014)	(0.006)	(0.006)
Dep Var$_{(t-1)}$		0.324^{*}	0.386^{*}		0.643^{***}	0.734^{***}
		(0.175)	(0.208)		(0.115)	(0.113)
Government	0.037^{***}			0.055^{***}		
	(0.009)			(0.010)		
Non-profit	0.017^{***}			0.028^{***}		
	(0.006)			(0.007)		
Urban	0.003			0.013		
	(0.008)			(0.010)		
Year Dummies	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1518	1518	1518	1518	1518	1518
Hospitals	323	323	323	323	323	323
R^2	0.44			0.40		
Instruments		50	40		50	40
AR(1) (p-value)		(0.000)	(0.002)		(0.000)	(0.000)
AR(2) (p-value)		(0.118)	(0.094)		(0.073)	(0.072)
Hansen test (p-value)		(0.200)	(0.082)		(0.571)	(0.702)
Diff.-in-Hansen test (p-value)		(0.225)	(0.465)		(0.163)	(0.645)

Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

A.13.2 Determinants of Practice Variation: Personal vs. Organizational

We have focused on either time-varying or time-invariant hospital-level characteristics to address potential determinants of practice variation. One other possibility is that the level of practice variation may vary mainly due to the physicians' idiosyncratic medical decision-making processes and behaviors, regardless of such organizational factors.

We thus tackle this issue using a descriptive approach. We track *big* attending physicians[5] who work at a minimum of two different hospitals and have 20 or more patients for each hospital in a year. If any physician works at more than two hospitals, we focus on the first two hospital with higher volumes. Our data contains 428 physician-years of such cases. We derive physician-level practice variation for each case. Risk-adjusted charge measures are again used in the calculation of practice variation to rule out patient-mix related variations. We posit that if the practice variation arising from personal behavior dominates that due to organizational factors (e.g., hospital characteristics discussed earlier and/or guidelines, pressures, and culture at the hospital), then practice variation of a focal physician across hospitals will be similar no matter what the hospital is.

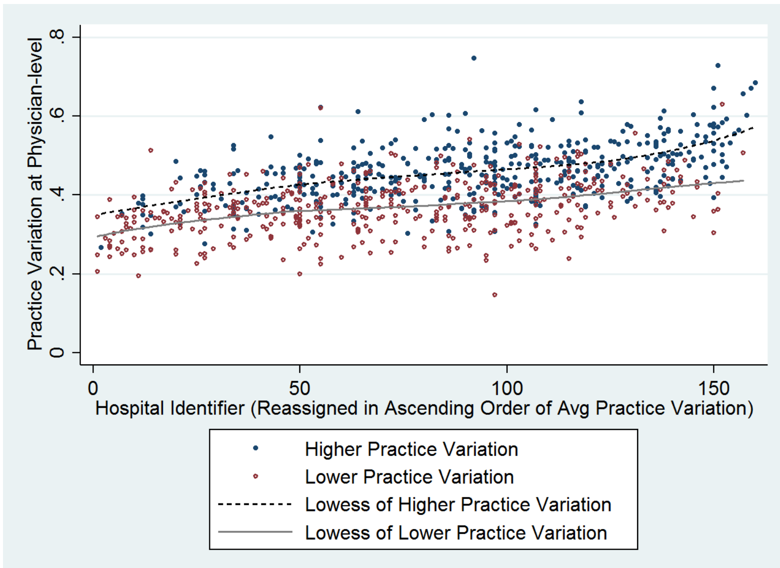
Figure A.5 summarizes our findings. The solid blue dots (the open red dots) represent higher (lower) practice variation between the two hospitals where a focal physician works. The dashed line (the solid) is the lowess[6] regression of the higher (lower) practice variation. The horizontal-axis is the hospital identifiers that are sorted in ascending order and reassigned in terms of average practice variation across hospitals. From the increasing pattern of the lowess either for higher or for lower practice variation, again we observe that the levels of practice variation varies across hospitals. More importantly and interestingly, we find a significant gap between the two lowess lines. Therefore, we claim that, on average, physician-level practice variation may depend on the hospitals that physicians are working at (although our insights are only applicable to the physicians with a high volume of patients). According to an unpaired t-test assuming unequal variance, the difference between higher and lower practice variations is significantly different from zero ($t-$

[5]Attending physicians have final responsibility for patient care, although some of the decisions can be made by others such as residents, medical students, or mid-level practitioners.

[6]Lowess: Locally Weighted Scatterplot Smoothing

stat $= -19.38$, $p < 0.001$). In other words, our analysis reveals that a significant portion of practice variation can be explained from organizational factors (even if we cannot fully ignore the effect of each physician's personal behavior in our measure of practice variation).

Figure A.5: Physician-level Practice Variation for Physicians who visit Multiple Hospitals

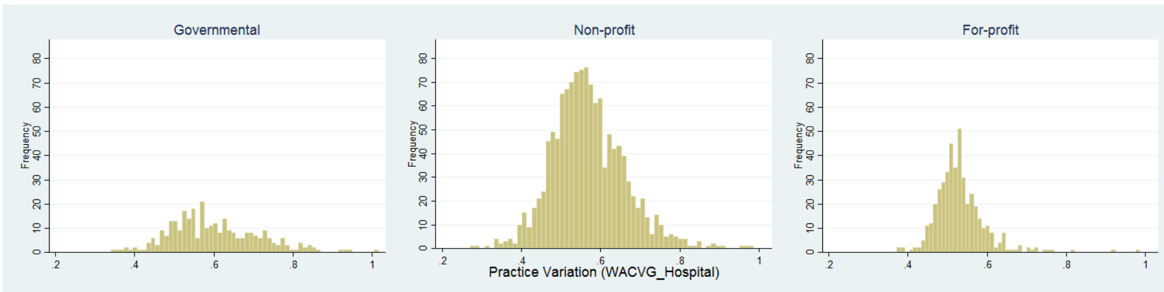


A.13.3 Practice Variation by Hospital Controls

In Appendix A.13, we discuss the determinants of practice variation. Although the magnitudes of coefficients are smaller for the system GMM estimates compared to OLS, the results are largely consistent. Time-invariant control variables such as governmental, non-profit, and urban disappear in the results of system GMM due to the first-difference in the process of estimation. In the static OLS estimates (i.e., Model (1) and (4) in Table A.17), we find that governmental or non-profit private hospital is associated with significantly higher level of practice variation, compared to for-profit private hospitals.

We again observe the patterns from the histograms shown in Figure A.6. Interestingly, we also find that the degree of dispersion in practice variation across governmental hospitals is higher than that of private hospitals. In other words, although the total number of governmental hospitals is much smaller than private hospitals, we observe a greater number of governmental hospitals than private hospitals with higher practice variation (e.g., $WACVG\_Hospital \geq 0.8$), indicating that variable practices are being used inside governmental hospitals.

Figure A.6: Histogram of Practice Variation by Hospital Control



A.14 Post-hoc Analysis: Practice Variation as Antecedent of Quality Measures

In this section, we address whether the process quality and experiential quality measures are observable measures of unobservable hospital-level decision-making processes. In other words, we examine the possibility of practice variation as an antecedent of these quality measures. We clarify from conceptual, practical, and statistical perspectives that the practice variation measure in our study is indeed capturing a distinct dimension of process variability, during a patient care episode, which is not necessarily related to either process quality or experiential quality. Thereby, we validate that the process and experiential quality measures satisfy the requirements to be considered as moderators.

Let us first discuss characteristics of the process quality measure, which are fundamentally different from the practice variation measure. First, as listed in Table A.6 in Appendix A.6, each metric in the process quality domain specifies a certain clinical protocol delivered to a patient cohort at a point of time (e.g., AMI-6: Patients given Beta blocker at arrival). Hence, the short-listed process metrics may not necessarily be affected by the overall process variability reflected in our practice variation measure. That is because, compared to the narrowed-boundary of the process quality metrics, our practice variation measure is designed to reflect all of the services and items provided during a patient care-episode using detailed medical charge information. As listed in M5 of Table A.18 and of Table A.19, we could not find any statistical evidence of practice variation as an antecedent of process quality. In these tables, we also provide the results for lagged rather than contemporaneous practice variation (M6 in each table) and OLS as an estimation strategy (M4 in each table). None of these settings provides significant evidence.

Second, in practice, process quality is often used to measure a hospital's "average" level of adherence to standard protocols, whereas our practice variation is more interested in capturing "dispersion" level of clinical services provided to a patient cohort.

Third, as requested by CMS, hospital managers are aware of the process quality, collect the process quality scores by themselves, and report them to CMS. We believe this is another important aspect that distinguishes our practice variation measure from the process quality measure. Our proposed approach for measuring practice variation did not exist during the period of data, and thus practitioners and healthcare managers were obviously not aware of their practice variation level. Hence, hospitals with high process quality possibly exhibit a wide spectrum of practice variation. In a simple two-by-two matrix, for example, each hospital can fall into one of the following combinations: (High PQ, High WACV), (High PQ, Low WACV), (Low PQ, High WACV), and (Low PQ, Low WACV), again supporting the insignificant results of the practice variation as an antecedent of process quality.

Similar arguments to those discussed above apply for the experiential quality as well. In addition, unlike process quality, which deals with attributes controllable by stakeholders inside a hospital, experiential quality is less controllable due to the patient engagement, also supporting the insignificant relationship between practice variation and experiential quality (as in M1 to M3 of Table A.18 and of Table A.19).

Table A.18: Practice Variation (WACVG) as Antecedent of Quality Measures

Dep Var	ExpQuality (EQ)			ProcQuality (PQ)		
Estimation	OLS	System GMM		OLS	System GMM	
	(M1)	(M2)	(M3)	(M4)	(M5)	(M6)
WACVG	0.033	-0.083		-0.052	-0.694	
	(0.051)	(0.346)		(0.255)	(0.560)	
WACVG$_{(t-1)}$			0.044			-1.112
			(0.716)			(0.721)
ProcQuality (PQ)	0.009	-0.030^{***}				
	(0.006)	(0.009)				
ExpQuality (EQ)				0.209	-1.809^{***}	
				(0.162)	(0.558)	
PQ$_{(t-1)}$			-0.027^{***}		0.719^{***}	0.743^{***}
			(0.008)		(0.064)	(0.059)
EQ$_{(t-1)}$		0.508^{***}	0.539^{***}			-1.714^{***}
		(0.073)	(0.091)			(0.469)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Year Dummies	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1518	1518	1518	1518	1518	1518
Hospitals	323	323	323	323	323	323
R^2	0.33			0.42		
Instruments		50	41		50	41
AR(1) (p-value)		(0.000)	(0.000)		(0.000)	(0.000)
AR(2) (p-value)		(0.130)	(0.098)		(0.592)	(0.288)
Hansen test (p-value)		(0.210)	(0.091)		(0.101)	(0.080)
Diff.-in-Hansen test (p-value)		(0.165)	(0.555)		(0.060)	(0.273)

Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.19: Practice Variation (WACVD) as Antecedent of Quality Measures

Dep Var	ExpQuality (EQ)			ProcQuality (PQ)		
Estimation	OLS	System GMM		OLS	System GMM	
	(M1)	(M2)	(M3)	(M4)	(M5)	(M6)
WACVD	0.093^{*}	0.156		-0.023	-0.804	
	(0.049)	(0.202)		(0.265)	(0.655)	
WACVD$_{(t-1)}$			0.540			-0.984
			(0.524)			(0.722)
ProcQuality (PQ)	0.009	-0.030^{***}				
	(0.005)	(0.009)				
ExpQuality (EQ)				0.208	-1.668^{***}	
				(0.162)	(0.533)	
PQ$_{(t-1)}$			-0.026^{***}		0.717^{***}	0.736^{***}
			(0.008)		(0.063)	(0.058)
EQ$_{(t-1)}$		0.496^{***}	0.486^{***}			-1.638^{***}
		(0.078)	(0.081)			(0.465)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Year Dummies	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1518	1518	1518	1518	1518	1518
Hospitals	323	323	323	323	323	323
R^2	0.44			0.40		
Instruments		50	41		50	41
AR(1) (p-value)		(0.000)	(0.000)		(0.000)	(0.000)
AR(2) (p-value)		(0.146)	(0.130)		(0.591)	(0.274)
Hansen test (p-value)		(0.072)	(0.158)		(0.151)	(0.220)
Diff.-in-Hansen test (p-value)		(0.524)	(0.548)		(0.084)	(0.091)

Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

A.15 Impact of Dispersion on Value

Most of the literature discussed in our study investigates association between explanatory variables and dependent variables which are both at "mean" levels. Several previous studies instead examine how the "dispersion" level of one metric impacts other variables. For example, Perdikaki et al. (2012) investigate how store traffic and labor relate to sales performance in the context of retail operations, where intra-day traffic variability is operationalized as traffic standard deviation divided by traffic mean. In the field of marketing, Luo et al. (2013) examine how brand rating dispersion impacts abnormal returns and idiosyncratic risk of a firm, where the brand dispersion is defined as a variance in brand ratings across consumers. Other examples include the relationship between customer satisfaction heterogeneity and shareholder value (Grewal et al., 2010) and applications of the judgment uncertainty and magnitude parameters (JUMP) model (e.g., Chandrashekaran et al., 2005).

Similarly, our research also constructs a variable for hospital-level practice variation, which is a measure of dispersion, and reports a descriptive analysis of how the practice variation correlates with operational performance. This is a valid research approach because there is potentially added information in variance that could inform process quality or experiential quality beyond the mean. Specifically, if the source of the variance is clinical practice discordance, variation would be a hurdle to reduce cost and to improve health outcomes, whereas if the source is stable heterogeneity in practice, variation could be a good thing (e.g., a sign of patient-centered practice), a bad thing (e.g., a sign of subjective or randomized practice that deviates from essential steps), or even a neutral signal.

APPENDIX B

SUPPLEMENT TO CHAPTER 3

B.1 DEA Formulation: BCC and output-oriented

The first DEA model, the Charnes-Cooper-Rhodes (CCR) model, developed by Charnes et al. (1978) is based on the assumption of constant return to scale (CRS). Later, Banker et al. (1984) enhance the CCR model and develop the Banker-Charnes-Cooper (BCC) model using the variable return to scale (VRS).

We introduce the formulation of BCC model. Since DEA is a non-parametric method, we don't need an explicit specification of the functional relationship between inputs and outputs (i.e., a production fuction)(Cherchye et al., 2000). These models can be distinguished by the envelopment surface and the orientation. CRS or VRS can be taken as the form of the envelopment surface. The surface of CRS is represented by a straight line that starts from the origin and passes through the first DMU that it meets as it approaches the observed values. Align with the meaning of CRS, the CRS surface model assume that an increase in inputs result in a proportional increase in outputs. On the other hand, the surface of VRS envelops the observations by connecting the outermost DMUs, including the one met by the CRS surface. The VRS model allows an increase in input values to result in a non-proportional increase of output levels. Decreasing returns to scale (DRS) occurs above the point where CRS and VRS meet, and increasing returns to scale (IRS) occurs below the point.

One other important characteristic of DEA models is orientation, which indicates the direction an inefficient DMU get closer to the efficient frontier. Input oriented represents a decrease in its input while keeping the same output level, whereas output oriented means an increase in its output levels while maintaining the same level of inputs. Input oriented DEA scores range between 0 and 1.0, and output oriented DEA scores range between 1.0 and infinity. However, it is same that both cases 1.0 is efficient.

The basic formulation can be transformed into LP by fixing the denominator to a constant value, e.g., 1.0, which can be interpreted as setting a constraint on the weights v_p. Now we get the LP model from the nonlinear original model which is included in Section 3.4.2. The subscript $o \in N$ denotes a focal DMU. Each HP, in turn, becomes a focal HP when its efficiency score is being computed.

$$\begin{aligned}
\underset{v,u}{\text{Maximize}} \quad & \sum_{q \in Q} u_q O_{qo} \\
\text{Subject to} \quad & \\
\sum_{p \in P} v_p I_{po} &= 1, && \text{(B.1)} \\
\sum_{q \in Q} u_q O_{qi} &\leq \sum_{p \in P} v_p I_{pi}, \forall i \in N && \text{(B.2)} \\
v_p, u_q &\geq 0
\end{aligned}$$

The problem above is solved once for each DMU to derive the relative efficiency scores. The dual formulation has been mentioned as being preferable from a computational point of view since typical primal models have many more rows than columns. The dual DEA model can be written as below.

$$\begin{aligned}
\underset{\lambda}{\text{Minimize}} \quad z_0 &= \theta_o \\
\text{Subject to} \quad & \\
\sum_{i \in N} \lambda_i O_{qi} &\geq O_{qo}, && \text{(B.3)} \\
\theta_o I_{po} &\geq \sum_{i \in N} \lambda_i I_{pi}, && \text{(B.4)} \\
\lambda_i &\geq 0
\end{aligned}$$

The model discussed above is the input oriented CCR model after the authors of Charnes et al. (1978). The input oriented BCC model can be obtained based on the dual, and adds a restriction on the λ's as in (B.5) that we call convexity constraint.

$$\sum_{i \in N} \lambda_i = 1 \tag{B.5}$$

This trasforms the model from being CRS to VRS. The scores from this model are called pure technical efficiency score as they eliminate scale-efficiency from the analysis (Vujcic and Jemric, 2001; Cooper et al., 1996). If we replace (B.3) and (B.4) with (B.6) and (B.7) given below respectively, then the problem becomes output oriented CCR model. Also, with addition of the convexity constraint (B.5), it becomes output oriented BCC model.

$$\sum_{i \in N} \lambda_i I_{pi} \leq I_{po} \tag{B.6}$$

$$\theta_o O_{qo} \leq \sum_{i \in N} \lambda_i O_{pi} \tag{B.7}$$

One difficult issue with every health application is whether to use DEA models that assume CRS or VRS. Researchers should address this question based on prior knowledge and logical inferences about the production context. While imaginative guesses are tolerable, it is unacceptable to pick a model to get âĂIJbetter lookingâĂİ DEA results (Cooper et al., 2011).

If the scope is defined in hospital-level, then the model introduced above can be directly used. However, if the scope is smaller than facility, let's say care process of each bundle, then it is meaningful to compare the relative efficiency level for the HPs who are interested in the bundle. Thus, we solve the following DEA model for each episode k in K with notations listed in Table B.1.

Table B.1: Notations used in CRS DEA model (Episode Level)

Parameters:	
I_{pi}^k	Value of input p parameter for HP i of episode k.
O_{qi}^k	Value of output q parameter for HP i of episode k.
Decision Variables:	
v_p^k	Weights for the inputs.
u_q^k	Weights for the outputs.

$$
\begin{aligned}
&\text{Maximize} && \frac{\sum_{q \in Q} u_q^k O_{qo}^k}{\sum_{p \in P} v_p^k I_{po}^k} \\
&\text{Subject to} && \\
&\frac{\sum_{q \in Q} u_q^k O_{qi}^k}{\sum_{p \in P} v_p^k I_{pi}^k} \leq 1, \forall i \in N, && \text{(B.8)} \\
&\frac{u_q^k}{\sum_{p \in P} v_p^k I_{pi}^k} \geq \epsilon, \forall q \in Q, && \text{(B.9)} \\
&\frac{v_p^k}{\sum_{p \in P} v_p^k I_{pi}^k} \geq \epsilon, \forall p \in p && \text{(B.10)}
\end{aligned}
$$

Constraints (B.8) ensure that no HP is more than 100 percents efficient. Constraint (B.9) and (B.10) ensure that each weight for the inputs and outputs is strictly positive since $\epsilon > 0$.

B.2 Package Combination Options under 4 Bundles

Table B.2: Package Combination Options under 4 Bundles

	DRG194	**DRG280**	**DRG291**	**DRG470**
Package 1	1	0	0	0
Package 2	0	1	0	0
Package 3	0	0	1	0
Package 4	0	0	0	1
Package 5	1	1	0	0
Package 6	1	0	1	0
Package 7	1	0	0	1
Package 8	0	1	1	0
Package 9	0	1	0	1
Package 10	0	0	1	1
Package 11	1	1	1	0
Package 12	1	1	0	1
Package 13	1	0	1	1
Package 14	0	1	1	1
Package 15	1	1	1	1

APPENDIX C

SUPPLEMENT TO CHAPTER 4

C.1 Table of Acronyms

Table C.1 contains a list of acronyms used in Chapter 4.

Table C.1: Table of Acronyms in Chapter 4

Acronym	Meaning
ASC	Ambulatory Surgery Center
HR	Holding Room
OR	Operating Room
PACU	Post Anesthesia Care Unit
CPT	Current Precedural Terminology
RCCP	Rough Cut Capacity Planning

C.2 Sequence of ASC Events

In Table C.2, we provide a detailed sequence of ASC events. We define (Operating Room In Time-Holding Room In Time) as HR duration and (Out Of Operating Room-Operating Room In Time) as OR duration. Similarly, we calculate (Patient Discharge-Out Of Operating Room) as PACU duration.

C.3 Details in Defining Patients Groups and Patients Sample Path

In this section, we provide supportive information on the way we define patient groups and patient sample path for the multi-specialty ASC. There are four admission type of patients, i.e., Early Morning Admission (EMA, who becomes inpatient), Inpatient (IP, who return back to their own ward after surgery), Observation (who stays overnight, but not necessarily), and Same-day Surgery (SAS, who most likely discharge shortly but also can stay overnight depending on recovery status). The majority of observation patients went through Phase 3 (85.4%, i.e., 164 out of 192 patients) while some SAS patients also went through Phase 3 (7.25%, i.e., 118 out of 1508). Note,

Table C.2: Sequence of ASC Events

Phase	Event Name	Process Order
HR	Admit Time	1
	Patient Ready for Preop	2
	Holding Room In Time	3
	HR Staff Done	4
	HR Anesthesia Ready	5
	HR Ready for OR	6
	Intraop Room Ready	7
	Anesthesia Start	8
OR	**Operating Room In Time**	9
	Induction	10
	Anesthesia Ready	11
	Position/Prep Start	12
	Time Out	13
	Procedure Start	14
	Procedure Conclusion Begun	15
	Incision Closed	16
	Procedure Stop	17
	OR Holding Start	18
PACU	**Out Of Operating Room**	19
	Anesthesia Finish	20
	Anticipated Out	21
	Ready For Attending	22

however, that not all the patients with Phase 3 are the overnight-stay. On average, SAS patients stay at Phase 3 shorter than Observation patients. IP patients are returned to their wards, so that average duration at Phase 3 is significantly shorter than other Admit Types. In sum, correctly classifying admission type is important (especially, Observation or SAS) in capacity planning, although it is impossible to be perfect.

As an additional patient characteristic, we may explicitly consider two types of anesthesia in the pre-op stage: regional block and general. In general, a patient with regional block anesthesia requires longer time in the Holding Room (HR, i.e., pre-operative stage) but shorter time for recovery in the Post Anesthesia Care Unit (PACU, i.e., post-operative stage) compared to a patient with general anesthesia. As surgery characteristics, we may consider ASA code[1] (codes 1-4), and surgical APGAR score for surgical complexity, in addition to service types and related Current Procedural Terminology (CPT) codes.

In our sample data, all patients stay in ASC 24 hours or less. Patients are classified as an

[1]ASA physical status classification system assesses the fitness of patients before surgery.

Table C.3: The number of clusters for each service type in low acuity ASC

Service Type	Service Type Name	# CPT Codes	# Obs.	# Clusters
1	OTOLARYNGOLOGY/H&N	98	523	4
2	OPHTHALMOLOGY	84	354	3
3	GENERAL ONCOLOGY SURGERY	50	330	4
4	ORTHOPEDIC	108	272	4
5	PLASTIC SURGERY	106	248	4
6	ORAL & MAXILLOFACIAL	32	79	4
7	GENERAL SURGERY	20	75	2
8	OTHERS	25	79	3
Total		523	1960	28

overnight-stay if they stay in phase 3 at least from 12 am to 7 am. Under this setting, there are patients whose phase 3 is more than 7 hours but still classified as non-overnight-stay.

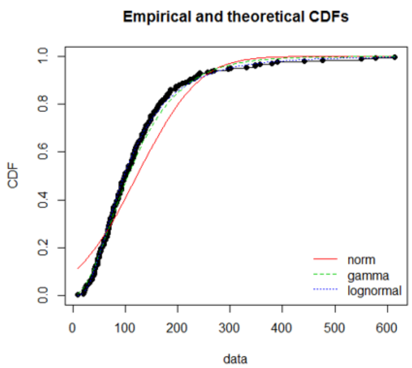
Initially, we relied on service type (e.g., orthopedic, plastic surgery) to fit the duration distribution at each stage. For example, for the patients classified as plastic surgery, we check the visualizations of distribution as in (a), (b), and (c) in Figure C.1 as well as Goodness-of-fit statistics (e.g., Kolmogorov-Smirnov, Carmer-von Mises, and Anderson-Darling statistics) and Goodness-of-fit criteria (e.g., AIC and BIC) to define a distribution for each service type and for each stage. We found that lognormal distribution fits nicely with the OR duration of entire plastic surgery patients. However, we realized that standard deviation is large as (d) in Figure C.1, meaning that the mean duration may not be very useful to run the deterministic model. Motivated by this result, we delve into CPT codes within a service type and conduct k-means cluster analysis to come up with patient clusters for each service type in terms of OR and PACU duration.

Table C.3 contains the number of clusters (i.e., k in k-means clustering). In determining the appropriate "k", we have compared several options listed in Table C.4[2] and select the reasonable size of clusters using visualization tools as in Figure C.2.

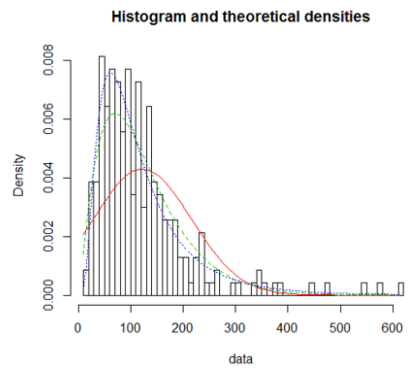
Afterwards, we provide fitted duration distribution for a similar group of clusters across the service types. The results are summarized in Table 4.4 and details are included in Table C.6 and C.7. We observe that average and standard deviation of HR duration are similar across service type

[2] http://stackoverflow.com/questions/15376075/cluster-analysis-in-r-determine-the-optimal-number-of-clusters

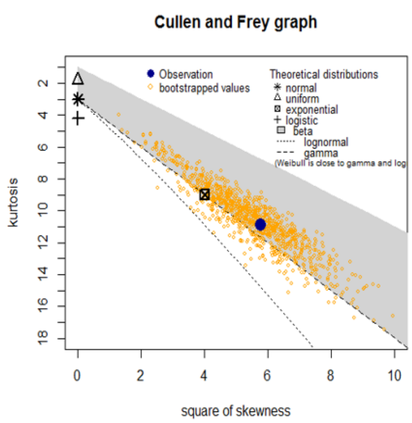
Figure C.1: Fitting the OR Duration Distribution of 234 Plastic Surgery Patients in Our Data



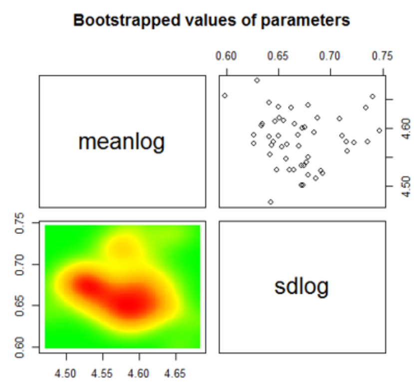
(a) Empirical and Theoretical CDFs



(b) Histogram and Theoretical Densities



(c) Cullen and Frey Graph

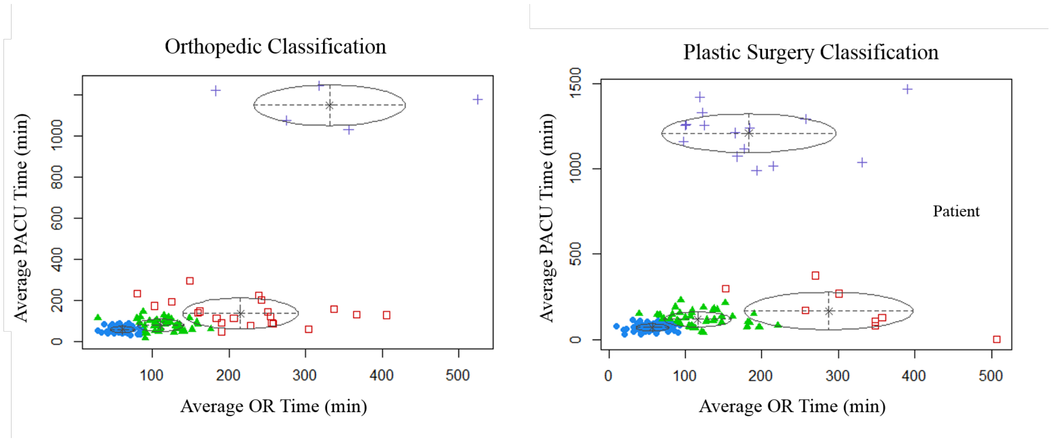


(d) Bootstrapped Values of Parameters

Table C.4: List of Options to Decide k in k-means Clustering

Option #	Description
1	Looked for a bend or elbow in the sum of squared error (SSE) scree plot. The location of the elbow in the resulting plot suggests a suitable number of clusters for the kmeans (See http://www.statmethods.net/advstats/cluster.html and http://www.mattpeeples.net/kmeans.html for more).
2	Two. Partitioned around medoids to estimate the number of clusters using the `pamk` function in the `fpc` package in R.
3	Calinsky criterion: Another approach to diagnosing how many clusters suit the data. We try 1 to 10 groups.
4	Determine the optimal model and number of clusters according to the Bayesian Information Criterion for expectation-maximization, initialized by hierarchical clustering for parameterized Gaussian mixture models using `mclust` package in R.
5	Affinity propagation (AP) clustering using `apcluster` in R (see http://dx.doi.org/10.1126/science.1136800).
6	Gap Statistic for Estimating the Number of Clusters with trying 2-10 clusters.

Figure C.2: Example of Clustering Results: Orthopedic (right) and Plastic Surgery(left)



and clusters, which again support the evidence that HR duration is independent of OR or PACU duration. We thus provide fitted duration distribution of HR duration for entire patients (see Table C.5), and of OR duration and PACU duration for each patient group (see Table C.6 and C.7).

Table C.5: Summary Statistics of HR Duration

ASC Type	Low Acuity
Cluster Type	All
Observations	1960
HR Obs.	1948
HR Duration Mean	136.33
HR Duration Std	67.62
HR Duration Median	121
HR Duration Min	16
HR Duration Max	621
HR Fitted Distribution	lognormal
Parameter 1: Median	4.81
Parameter 1: 95% CI	[4.79 4.82]
Parameter 2: Median	0.47
Parameter 2: 95% CI	[0.46 0.49]

To understand the variation in patient sample path, we plot the weekly time series by the seven patient cluster types as in Figure C.3[3]. We observe a significant variation in patient demand, for example, in week 3 there are 74 patients that could be classified as Cluster Type 1 while there are

[3]Out of 14-week data we have, we plot nine complete weeks. The other five weeks have either holidays or missing value as they are the beginning/ending point of the sample data.

Table C.6: Summary Statistics of OR and PACU Duration for each Patient Group (Part 1)

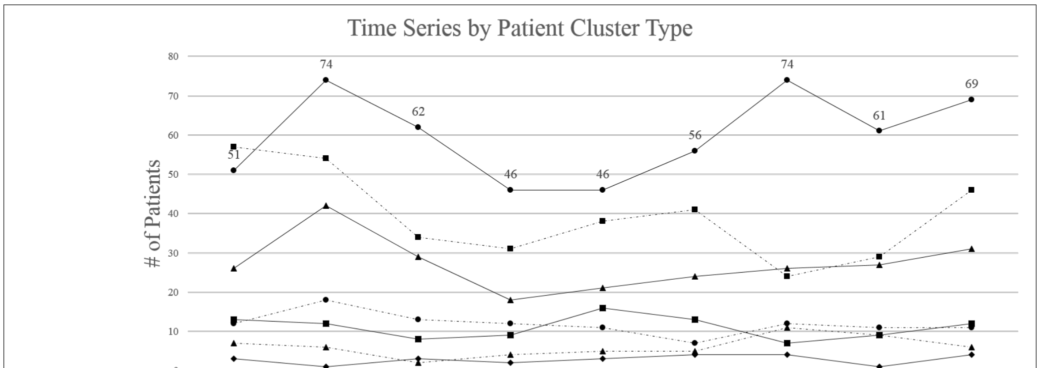
ASC Type	Low Acuity	Low Acuity	Low Acuity
Cluster Type	ShortOR-ShortPACU	ShortOR-ModeratePACU	ModerateOR-ShortPACU
Observations	735	674	176
OR Obs.	727	333	154
OR Duration Mean	68.43	77.91	137.73
OR Duration Std	34.58	34.73	57.74
OR Duration Median	63	71	134
OR Duration Min	10	12	34
OR Duration Max	305	217	361
OR Fitted Distribution	lognormal	lognormal	gamma
Parameter 1: Median	4.11	4.26	6.05
Parameter 1: 95% CI	[4.09 4.15]	[4.21 4.30]	[5.240 7.572]
Parameter 2: Median	0.49	0.45	0.04
Parameter 2: 95% CI	[0.46 0.51]	[0.43 0.48]	[0.037 0.055]
PACU Obs.	725	333	152
PACU Duration Mean	66.45	99.09	100.97
PACU Duration Std	38.46	69.94	63.50
PACU Duration Median	58	82	85
PACU Duration Min	13	21	36
PACU Duration Max	479	591	381
PACU Fitted Distribution	lognormal	lognormal	lognormal
Parameter 1: Median	4.09	4.45	4.49
Parameter 1: 95% CI	[4.05 4.12]	[4.40 4.49]	[4.42 4.55]
Parameter 2: Median	0.46	0.51	0.48
Parameter 2: 95% CI	[0.44 0.49]	[0.47 0.53]	[0.43 0.54]

Table C.7: Summary Statistics of OR and PACU Duration for each Patient Group (Part 2)

ASC Type	Low Acuity	Low Acuity	Low Acuity	Low Acuity
Cluster Type	ModerateOR-ModeratePACU	ModerateOR-OvernightPACU	LongOR-ModeratePACU	LongOR-OvernightPACU
Observations	512	45	108	141
OR Obs.	510	33	68	127
OR Duration Mean	144.82	166.27	280.29	225.25
OR Duration Std	56.41	81.58	85.09	87.31
OR Duration Median	135.5	164	264	214
OR Duration Min	44	59	99	32
OR Duration Max	525	390	521	536
OR Fitted Distribution	lognormal	lognormal	gamma	gamma
Parameter 1: Median	4.91	5.03	10.09	6.72
Parameter 1: 95% CI	[4.87 4.9]	[4.80 5.20]	[7.870 15.196]	[5.409 8.352]
Parameter 2: Median	0.38	0.45	0.04	0.03
Parameter 2: 95% CI	[0.36 0.4]	[0.34 0.56]	[0.028 0.058]	[0.024 0.037]
PACU Obs.	510	33	67	127
PACU Duration Mean	125.41	1155.60	173.97	1201.30
PACU Duration Std	65.59	173.25	87.97	151.75
PACU Duration Median	110	1175	147.00	1209
PACU Duration Min	23	827	60	865
PACU Duration Max	538	1467	455	1482
PACU Fitted Distribution	lognormal	weibull	lognormal	weibull
Parameter 1: Median	4.72	8.30	5.04	9.10
Parameter 1: 95% CI	[4.68 4.76]	[7 10]	[4.96 5.14]	[8 11]
Parameter 2: Median	0.46	1229.70	0.45	1267.20
Parameter 2: 95% CI	[0.44 0.49]	[1186 1279]	[0.37 0.55]	[1248 1287]

only 46 patients in week 8.

Figure C.3: Patient Flow at the Ambulatory Surgery Center



	Week2	Week3	Week7	Week8	Week9	Week10	Week11	Week12	Week13
Cluster Type 1	51	74	62	46	46	56	74	61	69
Cluster Type 2	26	42	29	18	21	24	26	27	31
Cluster Type 3	12	18	13	12	11	7	12	11	11
Cluster Type 4	57	54	34	31	38	41	24	29	46
Cluster Type 5	3	1	3	2	3	4	4	1	4
Cluster Type 6	7	6	2	4	5	5	11	9	6
Cluster Type 7	13	12	8	9	16	13	7	9	12

C.4 Formal Description of Heuristic BackwardASC

Let a_i be the time point when OR M_{1i} becomes idle $(i = 1, 2, \cdots, R_1)$ and b_i be the time point when PACU $M_{2,i}$ becomes idle $(i = 1, 2, \cdots, R_2)$. $\mathcal{G}$ stands for the ordered scheduling list of the patients and $\mathcal{I}$ represents the indices set of the patients not scheduled yet. $\mathcal{I}^o$ (respectively, $\mathcal{I}^{no}$) is the indices set of overnight-stay patients (respectively, non-overnight patients). We provide a formal description of Heuristic *BackwardASC* in Algorithm 1.

C.5 Illustration of Heuristics

Illustration of Algorithm *FFP1* (Upper Bound of ORs and PACUs)

Example 1: There are six patients, $\mathbf{P} = \{P_1, P_2, ..., P_6\}$, to be served in an ASC in a day. The durations at (OR, PACU) stages are $\{(9,3), (11,2), (7,4), (6,5), (4,4), (3,3)\}$ time units where 1 time unit is 0.5 hours. Note that we only consider the Non-overnight PACU (NOPACU) stay. Algorithm *FFP1* is a greedy algorithm that derives the upper bound of the number of ORs and

Algorithm 1 Heuristic *BackwardASC*

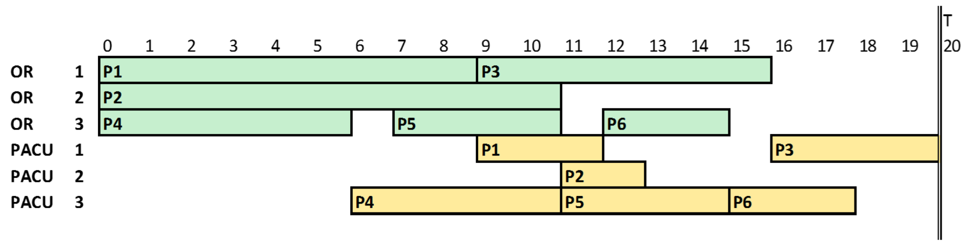
Input: R_1 ORs, $\{M_{1,1}, M_{1,2}, ..., M_{1,R_1}\}$, and R_2 PACUs, $\{M_{2,1}, M_{2,2}, ..., M_{2,R_2}\}$. $\mathcal{I} = \mathcal{I}^o \cup \mathcal{I}^{no} = \{1, 2, ..., I\}$, a set of patients to be scheduled where $\mathcal{I}^o$ is the set of overnight-stay patients and $\mathcal{I}^{no}$ is the set of non-overnight-stay patients. (p_{1i}, p_{2i}), the duration times of patient $j \in \mathcal{I}$ in (OR, PACU). $\mathcal{G} = \emptyset$, an ordered scheduling list of patients.
Step 0: $a_i \leftarrow 0, \forall i \in \{1, 2, \cdots R_1\}$ and $b_i \leftarrow 0, \forall i \in \{1, 2, \cdots, R_2\}$.
Step 1: $t_1 \leftarrow \min_i\{a_i\}$ and $k_1 \leftarrow \text{argmin}_i\{a_i\}$ (break ties arbitrarily).
$t_2 \leftarrow \min\{b_i\}$ and $k_2 \leftarrow \text{argmin}\{b_i\}$ (break ties arbitrarily). Let $t = \max\{0, t_2 - t_1\}$.
Step 2:
if $\mathcal{I}^o \neq \emptyset$ **then** $i \leftarrow \text{argmin}_{i \in \mathcal{I}^o}\{|p_{1i} - t|\}$ (break ties arbitrarily).
    $\mathcal{I}^o \leftarrow \mathcal{I}^o \setminus \{i\}$ and $\mathcal{G} \leftarrow \mathcal{G} \cup \{i\}$.
else$i \leftarrow \text{argmin}_{i \in \mathcal{I}^{no}}\{|p_{1i} - t|\}$ (break ties arbitrarily).
    $\mathcal{I}^{no} \leftarrow \mathcal{I}^{no} \setminus \{i\}$ and $\mathcal{G} \leftarrow \mathcal{G} \cup \{i\}$.
end if
Step 3:
if $p_{1i} \geq t$ **then** $a_{k_1} \leftarrow a_{k1} + p_{1i}$ and then $b_{k_2} \leftarrow a_{k_1} + p_{2i}$
else $a_{k_1} \leftarrow b_{k_2}$ and $b_{k_2} \leftarrow b_{k_2} + p_{2i}$.
end if
Step 4: If $\mathcal{I}^o = \mathcal{I}^{no} = \emptyset$, Stop. Otherwise, go to Step 1.
Output: $\mathcal{G}$, the ordered scheduling list of patients.

NOPACUs without overtime. Given list of patient $\mathbf{P}$, we first assign patient P_1 at time 0 to OR_1 and $PACU_1$ accordingly with no-wait constraint between the two stages. Second, when P_2 is assigned to the same OR and PACU, overtime, i.e., the duration beyond the regular hour T, is expected. Hence, we increase the number of OR and PACU by one and schedule at time 0. Next, we assign patient P_3 to the first set of OR and PACU as no overtime is expected. Conduct the same process until all patients are assigned. Thus, Algorithm *FFP1* generates three ORs and PACUs with the following daily schedule: P_1, P_3 are allocated to (OR_1, $PACU_1$) in this order; P_2 is allocated to ($OR_2, PACU_2$); and P_4, P_5, P_6 are allocated to (OR_3, $PACU_3$) in this order. See Figure C.4. Once we know the expected number of overnight-stay patients M, e.g., 2, then 3 ORs and 5 PACUs become the initial settings in *AdaptiveASC* algorithm.

Illustration of Heuristic *BackwardASC*

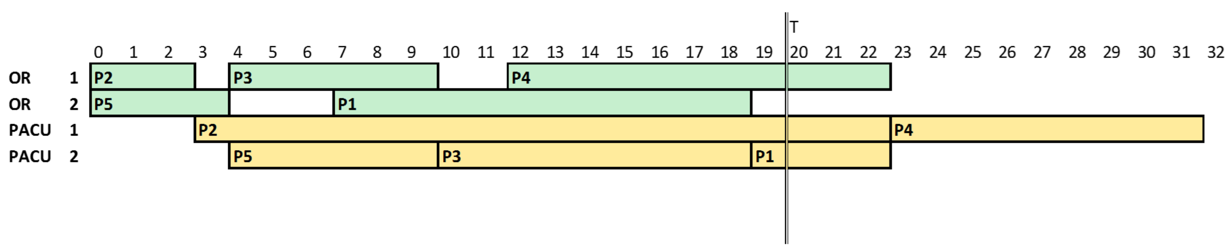
Example 2: There are five patients, $\mathbf{P} = \{P_1, P_2, ..., P_5\}$, to be served in an ASC in a day. The number of rooms at (OR, PACU) is (2,2). The durations at (OR, PACU) stages of the patients are $\{(12, 4), (3, 20), (7, 9), (11, 9), (4, 6)\}$ (in this example we assume $\mathcal{I}^o = \emptyset$, i.e., there is no overnight-stay patients). When the heuristic begins, $t_1 = t_2 = t = 0$, and hence, patient P_2 is

Figure C.4: The schedule obtained by Algorithm FFP1: Derived number of ORs and NOPACUs are (3,3)



allocated first to (OR_1, $PACU_1$). Second, we still have $t_1 = t_2 = t = 0$, and patient P_5 is assigned to (OR_2, $PACU_2$). Third, $t_1 = 3$, $t_2 = 10$, $t = 7$, and thus, patient P_3 is processed at (OR_1, $PACU_2$). Fourth, $t_1 = 4$, $t_2 = 19$, $t = 15$, patient P_1 is assigned next at (OR_2, $PACU_2$). Lastly, $t_1 = 10$, $t_2 = 23$, $t = 13$, and only patient P_4 can be processed next and we assign her to (OR_2, $PACU_1$). See Figure C.5.

Figure C.5: The schedule obtained by Heuristic BackwardASC: OR and PACU overtimes are 3 and time units, respectively



C.6 Sequence-based Formulation of Problem P1

While the formulation of Problem *P1* in Section 4.5.2.1 is based on the discretized time slots over planning horizon, the sequence-based version of Problem *P1* with both overtime and idle time in each stage can be formulated as follows (Additional notations are listed in Table C.8).

Table C.8: Additional notations for the sequence-based formulation of *P1*

Parameters:	
PRT_i	Earliest available time of patient i.
Decision Variables:	
$x_{ii's}$	If $x_{ii's} = 1$, patient i' is served right after patient i in stage s.
ts_{isr}	Time at which patient i enters room r in stage s.
tf_{isr}	Time at which patient i leaves room r in stage s.
y_{isr}	$y_{isr} = 1$ if $ts_{isr} > 0$.
CT_{sr}	Makespan of room r in stage s.
OT_{sr}	Overtime of room r in stage s.
IT_{sr}	Idle time of room r in stage s.

The objective function is to minimize the sum of idle time and overtime:

$$\min \sum_{s\in\mathcal{S}} \sum_{r\in\mathcal{R}_s} \left[C_s^d IT_{sr} + C_s^o OT_{sr} \right] + \sum_{s\in\mathcal{S}} C_s^e R_s \tag{C.1}$$

Each patient i can enter the stage 1 after at least his earliest available time:

$$ts_{i1b} \geq PRT_i \quad \forall i \in \mathcal{I},\ \forall r \in R_1, \tag{C.2}$$

The time when patient i enters stage $s \geq 2$ is the sum of previous stage's entering time and the duration in stage $s-1$:

$$\sum_{b\in\mathcal{R}_s} ts_{isr} = \sum_{r\in B(s-1)} ts_{isr} + p_{i,s-1} \qquad \forall i \in \mathcal{I},\ \forall s \in \mathcal{S}/\{1\}, \tag{C.3}$$

Let $I_{is} = \{i' \in \mathcal{I} \mid \text{patient } i' \text{ can follow directly after patient } i \text{ on stage } s\}$. A single patient i' can be served after patient i, vice versa:

$$\sum_{i'\in\mathcal{I}_{is}} x_{ii's} \leq 1 \qquad \forall i \in \mathcal{I},\ \forall s \in \mathcal{S}, \tag{C.4}$$

$$\sum_{i\in\mathcal{I}_{i's}} x_{i'is} \leq 1 \qquad \forall i' \in \mathcal{I},\ \forall s \in \mathcal{S}, \tag{C.5}$$

$$x_{ii's} + x_{i'is} \leq 1 \qquad \forall i \in \mathcal{I},\ \forall i' \in \mathcal{I},\ \forall s \in \mathcal{S}, \tag{C.6}$$

As long as $I \geq R_s, \forall s \in \mathcal{S}$, we can assume that all rooms will process at least one patient with no loss of generality. In this case, there will be exactly R_s patients with no followers:

$$\sum_{i \in \mathcal{I}} \sum_{i' \in \mathcal{I}_{is}} x_{ii's} = I - \mathcal{R}_s \qquad \forall i \in \mathcal{I},\ \forall r \in R_1, \tag{C.7}$$

Patient i' cannot enter stage s while patient i' is being served in the stage s:

$$\sum_{b \in \mathcal{R}_s} ts_{isr} + p_{is} \leq \sum_{r \in \mathcal{R}_s} ts_{i'sb} + M(1 - x_{ii's}) \qquad \forall i \in \mathcal{I},\ \forall s \in \mathcal{S}, \tag{C.8}$$

Similarly, the entering time of patient i in stage $(s+1)$ is at least as fast as the entering time of patient i':

$$\sum_{b \in \mathcal{R}_s} ts_{i(s+1)b} \leq \sum_{r \in \mathcal{R}_s} ts_{i'sb} + M(1 - x_{ii's}) \qquad \forall i \in \mathcal{I},\ \forall s \in \mathcal{S}/\{2\}, \tag{C.9}$$

The entering time of patient i in stage s should be greater than the sum of durations of previous patient i':

$$\sum_{b \in \mathcal{R}_s} ts_{isr} \geq \sum_{i' \in \mathcal{I}_{is}} x_{i'is}(PRT_{i'} + \sum_{s' \leq s} P_{i's'}) \qquad \forall i \in \mathcal{I},\ \forall s \in \mathcal{S}, \tag{C.10}$$

If $ts_{i'sb} > ts_{isr}$, then $ts_{i'sb} \geq tf_{isr}, \forall i,\ s,\ b$. This if-statement can be written as follows:

$$0 \leq (ts_{isr} - ts_{i'sb}) + M \cdot d_{i'isr} \qquad \forall i, i' \in \mathcal{I},\ \forall r \in \mathcal{R}_s,\ \forall s \in \mathcal{S}, \tag{C.11}$$

$$tf_{isr} - ts_{i'sb} \leq M \cdot (1 - d_{i'isr}) \qquad \forall i, i' \in \mathcal{I},\ \forall r \in \mathcal{R}_s,\ \forall s \in \mathcal{S}, \tag{C.12}$$

where $d_{i'isr} \in \{0, 1\}, \forall i, i' \in \mathcal{I},\ \forall r \in \mathcal{R}_s,\ \forall s \in \mathcal{S}$.

We write the definition of y_{isr}, i.e., $y_{isr} \geq 1$ if $ts_{isr} > 0$ as the following set of constraints:

$$0 \leq (y_{isr} - 1) + M \cdot z_{isr} \quad \forall i \in \mathcal{I},\ \forall r \in \mathcal{R}_s,\ \forall s \in \mathcal{S}, \tag{C.13}$$

$$t_{isr} \leq M \cdot (1 - z_{isr}) \quad \forall i \in \mathcal{I},\ \forall r \in \mathcal{R}_s,\ \forall s \in \mathcal{S}, \tag{C.14}$$

where $z_{isr} \in \{0, 1\}$, $\forall i,\ s,\ b$. Each patient i in each stage s can only occupy a single room:

$$\sum_{r \in \mathcal{R}_s} y_{isr} = 1 \quad \forall i \in \mathcal{I},\ \forall s \in \mathcal{S}, \tag{C.15}$$

Patient i in room r and stage s leaves after the duration in stage s:

$$tf_{isr} = ts_{isr} + p_{is} y_{isr} \quad \forall i \in \mathcal{I},\ \forall r \in \mathcal{R}_s,\ \forall s \in \mathcal{S}, \tag{C.16}$$

The makespan of room r in stage s is the maximum completion time across all patients:

$$CT_{sr} = \max_{i \in \mathcal{I}} \{tf_{isr}\} \quad \forall r \in \mathcal{R}_s,\ \forall s \in \mathcal{S}, \tag{C.17}$$

Let $h_{isr} \in \{0, 1\}$, $\forall i,\ s,\ b$. Also, let the upper bound of tf_{isr} be UB_{isr}, $\forall i,\ s,\ b$. Then, the constraint (C.17) can be rewritten as the following set of constraints:

$$0 \leq t_{isr} \leq UB_{isr} \quad \forall i \in \mathcal{I},\ \forall r \in \mathcal{R}_s,\ \forall s \in \mathcal{S}, \tag{C.18}$$

$$CT_{sr} \geq tf_{isr} \quad \forall i \in \mathcal{I},\ \forall r \in \mathcal{R}_s,\ \forall s \in \mathcal{S}, \tag{C.19}$$

$$CT_{sr} \leq tf_{isr} + UB_{isr}(1 - h_{isr}) \quad \forall i \in \mathcal{I},\ \forall r \in \mathcal{R}_s,\ \forall s \in \mathcal{S}, \tag{C.20}$$

$$\sum_{i \in \mathcal{I}} h_{isr} = 1 \quad \forall r \in \mathcal{R}_s,\ \forall s \in \mathcal{S}, \tag{C.21}$$

Overtime of room r in stage s is defined as $\max\{CT_s b - T, 0\}$, $\forall b \in \mathcal{R}_s,\ s \in \mathcal{S}$:

$$OT_{sr} \geq CT_{sr} - T \quad \forall r \in \mathcal{R}_s,\ s \in \mathcal{S} \tag{C.22}$$

In this sequence-based formulation, the idle time IT_{isr} is the most intricate variable to capture properly:

$$IT_{sr} \geq T - \sum_{i \in \mathcal{I}} \max\left[\min\{T,\ tf_{isr}\} - ts_{isr}, 0\right] \qquad \forall r \in \mathcal{R}_s,\ \forall s \in \mathcal{S}. \tag{C.23}$$

Collectively, the sequence-based of Problem *P1* can be formulated as:

$$\begin{aligned} \min \quad & \sum_{s \in \mathcal{S}} \sum_{r \in \mathcal{R}_s} \left[C_s^d IT_{sr} + C_s^o OT_{sr}\right] + \sum_{s \in \mathcal{S}} C_s^e R_s \\ \text{subject to} \quad & \text{Constraints (C.2) - (C.23)} \end{aligned}$$

C.7 Alternative Heuristic to Solve Problem P1 when $R_1 = R_2 = R$

We propose an alternative heuristic, *BackwardASC*$_{\mathbf{Alt}}$, which is devised based on the heuristic proposed by Sriskandarajah (1993) for problem $F_2(m_1, m_2)|$no-wait, $m_1 = m_2 = m \geq 2|C_{max}$. We begin with the m parallel processing centers that comprises m ORs and m PACUs (thus, each processing center is a simple two-machine no-wait flow shop). Heuristic *BackwardASC*$_{\mathbf{Alt}}$ assigns patients to the m parallel processing centers and derive a sequence of the patients assigned to each processing center. Specifically, the heuristic calculates the total treatment time of each patient by summing the processing times at both intra- and post-op rooms. Subsequently, the heuristic evenly allocates the total sum of treatment times to m parallel processing centers using Multi-fit algorithm (Friesen, 1984). Afterwards, the sequence of patients is determined in each of the m two-machine flow shops by the algorithm proposed by Gilmore and Gomory (1964).

To apply *BackwardASC*$_{\mathbf{Alt}}$, we use the patient sample path and duration information as listed in Table 4.6. *BackwardASC* algorithm calculates the total cost when the number of rooms for OR and PACU are and $M + |\mathcal{I}^o_{M_j}|$, respectively, where $|\mathcal{I}^o_{M_j}|$ is the maximum number of overnight-stay patients in a day across weekdays[5]. We iteratively run *BackwardASC*$_{\mathbf{Alt}}$ for multiple cases by changing the size of $\mathcal{R}_s$ and then compare the results described in Step 4 of the algorithm.

[5]In our scenario, $\mathcal{I}^o_{M_j} = \max\{4, 2, 2, 3, 2\} = 4$ where the numbers are the rounded sum of average *ModerateOR-OvernightPACU* patients and average *ModerateOR-OvernightPACU* patients listed in Table 4.5

Algorithm 2 Heuristic *BackwardASC*$_{\mathbf{Alt}}$

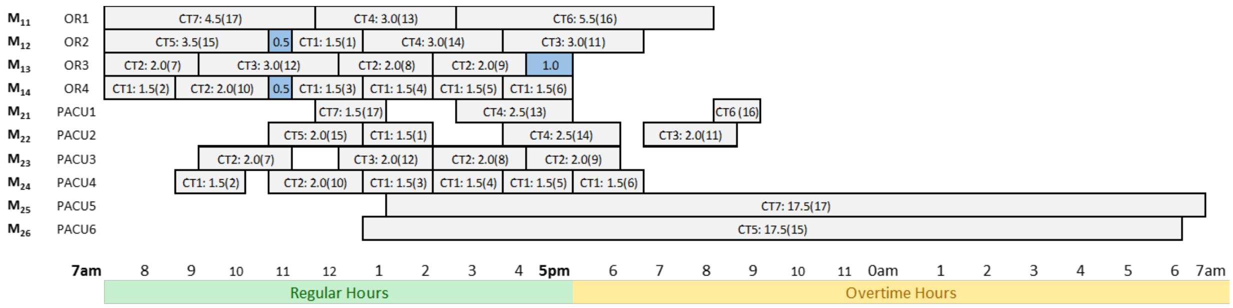
Input: R ORs, $\{M_{1,1}, M_{1,2}, ..., M_{1,R}\}$, and R PACUs, $\{M_{2,1}, M_{2,2}, ..., M_{2,R}\}$. $\mathcal{I} = \mathcal{I}^o \cup \mathcal{I}^{no} = \{1, 2, ..., I\}$, a set of patients to be scheduled where $\mathcal{I}^o$ is the set of overnight-stay patients and $\mathcal{I}^{no}$ is the set of non-overnight-stay patients. (p_{1i}, p_{2i}), the duration times of patient $i \in \mathcal{I}$ in (OR, PACU).
Step 1: Partition the problem into m flowshop problems each having two centers (i.e., OR and PACU) with each center having exactly one room.
Step 2: A virtual parallel machine shop with m rooms is formed by considering each flow shop $\{M_{1,j}, M_{2,j}\}$, as a single room M_j. Using **Multi-fit algorithm**, schedule I patients (of processing time requirements $p_1, p_2, \cdots, p_I$) in the parallel machine shop. In this way, patients are allocated to m flowshops. Three different approaches of Multi-fit available as below:
∘ Multi-fit approach 1: p_i based on sum of OR and non-overnight PACU durations, i.e., $p_i = p_{1i} + p_{2i}$.
∘ Multi-fit approach 2: p_i based on OR duration only, i.e., $p_i = p_{1i}$.
∘ Multi-fit approach 3: p_i based on non-overnight PACU duration only, i.e., $p_i = p_{2i}$.
Step 3: Let $\mathcal{I}_{M_j} = \mathcal{I}^o_{M_j} \cup \mathcal{I}^{no}_{M_j}$ be a set of patients for each flowshop M_j, where $\mathcal{I}^o_{M_j}$ is a set of overnight patients and $\mathcal{I}^{no}_{M_j}$ is a set of non-overnight patients.
if $\mathcal{I}^o_{M_j} = \emptyset$ **then**
The algorithm proposed by Gilmore and Gomory (1964)[4] optimally solve the flowshop M_j.
else
Apply Gilmore and Gomory (1964) algorithm to $\mathcal{I}^o_{M_j}$ and $\mathcal{I}^{no}_{M_j}$ separately.
Sequence the optimal order of jobs from $\mathcal{I}^o_{M_j}$ first and then from $\mathcal{I}^{no}_{M_j}$.
Add $|\mathcal{I}^o_{M_j}|$ number of PACUs for overnight patients.
end if
Step 4: Let $\mathcal{I}^*_{M_j}$ be the optimal sequence in the flowshop M_j. Calculate makespan, overtime cost at each stage based on predefined regular hours (e.g., 7am to 5pm for ORs).
Output: The ordered scheduling list of patients in each flowshop M_j.

• **Example of *BackwardASC*$_{\mathbf{Alt}}$:** Let $I = 17$, $R = 4$. The durations of the patients are listed in Table C.9. Remind that *BackwardASC*$_{\mathbf{Alt}}$ begins with the case $R_1 = R_2 = 4$ and finalize with the increased number of R_2 depending on the number of overnight-stay patients. There are two overnight-stay patients in this example (i.e., $i = 15,\ 17$), thus R_2 becomes 6. Figure C.6 illustrates the scheduling results of Heuristic *BackwardASC*$_{\mathbf{Alt}}$.

Table C.9: The durations of the patients for the example.

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
p_{1i}	1.5	1.5	1.5	1.5	1.5	1.5	2.0	2.0	2.0	2.0	3.0	3.0	3.0	3.0	3.5	5.5	4.5
p_{2i}	1.5	1.5	1.5	1.5	1.5	1.5	2.0	2.0	2.0	2.0	2.0	2.0	2.5	2.5	2.0	1.0	1.5
p_{2i}^{over}	-	-	-	-	-	-	-	-	-	-	-	-	-	-	17.5	-	17.5

Figure C.6: Patient scheduling generated by BackwardASC (with multifit on OR+PACU)



Notes: Patient schedules are based on the example in Table C.9 (initial $R_1 = R_2 = 4$)). The value in each cell indicates "Cluster Type: Duration(Patient i)." $C_{\max} = 14$, $LB_{C_{max}} = 11.63$, $Ratio = 1.2038$.

C.8 Forward Capacity Planning Approach

In the backward capacity planning approach, we have scheduled patients in ORs and PACUs using Heuristic *BackwardASC*, and then determined the required number of HRs using *Minimum Cost Flow (MCF)* model that preserves the scheduling results obtained from *BackwardASC*. This approach is called *backward ASC scheduling* algorithm because we first schedule OR and PACU, which are both resource intensive compared to HR. In this section, given the number of HRs (i.e., the output of *MCF*), we develop a *forward ASC scheduling* algorithm that particularly resembles the practice in which patients are scheduled from the beginning of the process. We compare the performances of both approaches.

In the forward approach, (1) we use Heuristic *ForwardASC* that similarly works as *BackwardASC* but schedules patients in the number of HR and ORs that are determined in the backward approach. (2) We next develop heuristic *NumPACU* to decide the minimal number of PACUs that preserve the patient schedule in the HR and ORs.

C.8.1 Heuristic ForwardASC

Refer to Heuristic *BackwardASC* and adjust it to apply in HR and OR stages rather than OR and PACU Stages.

C.8.2 Obtain the Number of PACUs

Similar to the *MCF* model in the backward approach, we can derive the number of PACUs to preserve the scheduling results obtained from *ForwardASC*. Based on the obtained starting time and ending time at HR and OR stages for n patients, we determine the number of PACUs that preserves the scheduling of patients in HRs and ORs.

C.8.3 Comparison of Backward Planning Approach to Forward Approach

To evaluate the average case performance of the backward and forward heuristics, we design computational experiments, compare the effectiveness of the heuristics under various conditions, and discuss our findings. In this section, we consider both overtime and idle time in calculating total cost.

We begin with the case of deterministic patient sample path and deterministic duration at each stage. Table C.10 compares the backward and forward capacity planning approaches regarding the total costs and their lower bound results as well as the derived number of rooms in HR and PACU, for each given number of OR. When we run Heuristic *BackwardASC*$_{\mathbf{Alt}}$, the multifit algorithm is applied to the sum of OR and PACU durations. Monday's patient sample path, for example, the capacity of $[HR, OR, PACU] = [18, 10, 14]$ (respectively, $[HR, OR, PACU] = [20, 11, 14]$) reveals the lowest total cost according to the backward approach (respectively, the forward approach). Across the patient sample paths for all five weekdays, the capacity of $[HR, OR, PACU] = [15, 9, 11.6]$ (respectively, $[HR, OR, PACU] = [19, 10, 14]$) reveals the lowest total cost for the backward approach (respectively, the forward approach). Under the objective of cost minimization, we observe that the backward capacity planning approach requires smaller numbers of rooms rather than the forward approach to schedule the same list of patients.

Uncertainty exists in daily patient sample path (as listed in Table 4.5), for example, the average number of *ShortOR-ShortPACU* patients (i.e., patient group 1) on Monday is 11.45 with standard deviation 3.33. We thus perform the simulation with instance size 1000 to reflect the stochastic patient sample path on capacity planning and list the results in Table C.11. Compared to the deterministic patient sample path in Table C.10, the smallest average total cost of backward approach (respectively, forward approach) has increased from 76.8 to 91.2, i.e., 18.8% (respectively, 81.5 to 98.4, i.e., 20.7%). Uncertainty in either patient-mix or duration of patient stays incurs additional costs, implying the benefits of simple patient-mix and accurate duration estimates in capacity planning. Across all five weekdays, the capacity of $[HR, OR, PACU] = [15.3, 9, 11.8]$ (respectively, $[HR, OR, PACU] = [18.5, 11, 12.9]$) reveals the lowest total cost for the backward approach (respectively, the forward approach).

In Figure C.7, we compare the results of computational experiments with stochastic Monday patient sample path for backward and forward approaches. According to the subfigure (b), as the number of ORs increases, the number of backward PACUs increases unlike the number of forward PACUs that almost remains as the same. The backward approach is more cost-efficient than forward approach, particularly when the number of ORs are smaller than or equal to the optimal (as in subfigure (g)) where the total cost is defined as a sum of the costs related to OR idle time, OR overtime, and PACU over-staffing time cost. That is because the backward approach has patient schedules with relatively smaller OR overtime (as in subfigure (e)) and PACU over-staffing time (as in subfigure (f)), even if the amount of OR idle time is a bit higher (as in subfigure (d)) than the forward approach.

Insight 1. *Backward planning approach is more cost-efficient than forward approach, particularly when the number of ORs are small. That is because the backward approach reveals less OR overtime and PACU overtime than the forward approach, even if OR idle time of the backward approach increases a bit.*

Table C.10: The Results of Backward vs. Forward ASC Capacity Planning: Deterministic Case

		Backward ASC Capacity Planning								Forward ASC Capacity Planning							
	OR	6	7	8	9	10	11	12	13	6	7	8	9	10	11	12	13
Monday	HR	10	12	13	15	18	20	19	19	10	12	13	15	18	20	19	19
	PACU	10	11	12	13	14	15	16	17	13	14	10	11	12	14	12	13
	Total Cost	160.3	128.3	113.0	102.8	**101.0**	107.3	123.3	132.8	225.0	192.5	152.3	130.3	113.3	**111.3**	121.5	130.3
	Cost/LB	3.6	3.6	4.2	5.6	10.1	7.4	5.2	4.1	7.4	9.0	3.5	3.7	4.2	7.7	4.5	4.0
Tuesday	HR	10	12	12	15	19	20	18	18	10	12	12	15	19	20	18	18
	PACU	8	9	10	11	12	13	14	15	13	13	11	13	15	17	12	13
	Total Cost	104.3	85.3	**61.8**	62.8	71.5	85.0	100.0	110.0	171.0	138.0	105.0	80.3	**69.0**	76.0	86.5	96.5
	Cost/LB	3.8	4.5	5.9	20.9	6.0	4.0	3.3	2.8	7.1	9.2	17.5	5.4	2.2	1.6	2.9	2.5
Wednesday	HR	11	14	15	14	20	21	18	19	11	14	15	14	20	21	18	19
	PACU	8	9	10	11	12	13	14	15	14	13	14	12	15	17	12	13
	Total Cost	106.3	75.8	74.8	**66.0**	73.8	80.0	95.0	110.0	175.3	133.0	107.3	87.8	**65.8**	70	81.5	90.0
	Cost/LB	3.4	3.3	5.2	11.0	6.1	3.8	3.2	2.8	7.3	8.9	5.5	29.2	2.3	1.6	2.7	2.3
Thursday	HR	11	13	14	14	19	20	17	18	11	13	14	14	19	20	17	18
	PACU	9	10	11	12	13	14	15	16	13	12	12	10	12	11	12	13
	Total Cost	141.3	104.5	95.3	**86.5**	86.5	91.8	106.8	117.8	204.3	166.3	138.0	111.5	**90.8**	92.0	104.3	114.0
	Cost/LB	3.9	3.8	5.0	8.2	13.3	5.9	4.4	3.5	6.9	8.1	12.0	4.1	8.6	4.8	4.3	3.4
Friday	HR	10	13	13	17	19	16	16	19	10	13	13	17	19	16	16	19
	PACU	8	9	10	11	12	13	14	15	12	13	12	12	16	11	12	13
	Total Cost	81.3	63.0	59.5	**66.0**	79.0	91.0	91.0	121.0	139.0	106.3	81.8	**63.0**	68.5	81.8	90.8	98
	Cost/LB	3.5	4.2	9.2	6.6	4.2	3.3	2.5	2.6	8.2	5.6	7.8	6.0	1.5	2.9	2.5	2.1
Avg HR		10.4	12.8	13.4	15	19	19.4	17.6	18.6	10.4	12.8	13.4	15	19	19.4	17.6	18.6
Avg PACU		8.6	9.6	10.6	11.6	12.6	13.6	14.6	15.6	13.0	13.0	11.8	11.6	14.0	14.0	12.0	13.0
Avg Total Cost		118.7	91.4	80.9	**76.8**	82.4	91.0	103.2	118.3	182.9	147.2	116.9	94.6	**81.5**	86.2	96.9	105.8
Avg Cost/LB Ratio		3.6	3.9	5.9	10.5	7.9	4.9	3.7	3.2	7.4	8.1	9.2	9.7	3.8	3.7	3.4	2.9

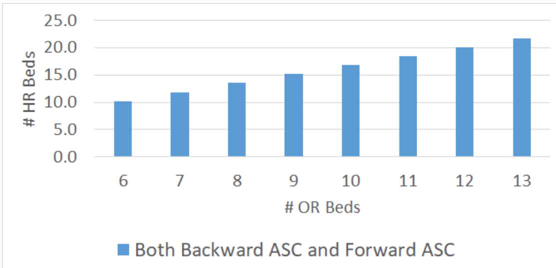
Notes: Deterministic Patient Sample Path and Deterministic Durations, OR idle cost $C^d = 1.5$, OR overtime cost $C_1^o = 2$, PACU over-staffing cost $C_2^o = 1$, and $6 \leq R_1 = R_2 \leq 13$

Table C.11: The Results of Backward vs. Forward ASC Capacity Planning: Stochastic Case

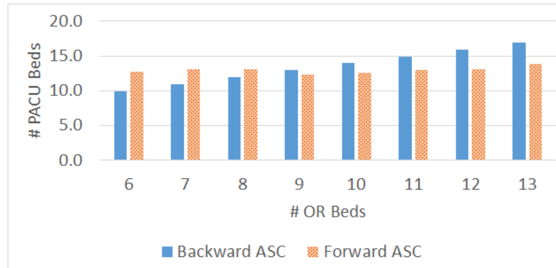
		Backward ASC Capacity Planning								Forward ASC Capacity Planning							
	OR	6	7	8	9	10	11	12	13	6	7	8	9	10	11	12	13
Monday	HR	10.2	11.9	13.6	15.2	16.9	18.5	20.1	21.7	10.2	11.9	13.6	15.2	16.9	18.5	20.1	21.7
	PACU	10.0	11.0	12.0	13.0	14.0	15.0	16.0	17.0	12.7	13.1	13.2	12.4	12.6	13.0	13.1	13.8
	Total Cost	164.8	141.2	124.6	115.8	**113.2**	116.1	123.1	133.1	225.7	192.6	164.2	142.4	127.8	**120.8**	121.3	124.3
	Cost/LB	3.6	3.8	4.4	5.8	7.9	7.1	5.1	4.0	6.5	6.5	6.0	4.9	4.8	4.9	4.8	3.8
Tuesday	HR	10.2	11.9	13.6	15.4	17.0	18.5	20.0	21.7	10.2	11.9	13.6	15.4	17.0	18.5	20.0	21.7
	PACU	8.2	9.2	10.2	11.2	12.2	13.2	14.2	15.2	12.0	12.7	13.2	12.9	13.2	13.8	13.7	14.3
	Total Cost	119.9	100.4	88.3	**83.7**	85.8	91.9	100.5	112.0	181.9	152.1	127.0	108.8	96.9	**91.9**	93.9	98.1
	Cost/LB	3.9	4.5	5.9	7.0	5.2	3.6	2.9	2.6	6.9	6.7	5.8	5.6	4.3	3.1	2.7	2.3
Wednesday	HR	10.3	12.1	13.7	15.5	17.1	18.7	20.1	21.3	10.3	12.1	13.7	15.5	17.1	18.7	20.1	21.3
	PACU	8.5	9.5	10.5	11.5	12.5	13.5	14.5	15.5	12.2	12.8	13.0	13.3	13.0	13.4	13.5	13.8
	Total Cost	124.8	103.6	90.2	**85.8**	87.5	93.8	103.0	114.4	188.5	158.7	132.3	111.7	99.8	**94.2**	95.7	100.8
	Cost/LB	4.1	4.8	6.6	8.4	5.8	4.0	3.2	2.8	7.2	6.9	6.2	5.3	4.7	3.5	2.9	2.5
Thursday	HR	9.9	11.6	13.3	14.9	16.7	18.2	19.4	20.4	9.9	11.6	13.3	14.9	16.7	18.2	19.4	20.4
	PACU	9.0	10.0	11.0	12.0	13.0	14.0	15.0	16.0	11.4	11.8	11.3	11.4	11.7	11.8	12.3	13.1
	Total Cost	141.1	118.3	103.8	96.5	**96.4**	100.7	108.7	118.5	197.1	166.7	140.5	120.6	108.5	**104.5**	106.3	111.4
	Cost/LB	3.7	4.0	5.0	7.3	8.6	5.8	4.2	3.4	6.1	6.3	5.8	5.8	5.7	5.7	4.3	3.3
Friday	HR	10.3	12.0	13.8	15.4	16.9	18.3	19.7	21.3	10.3	12.0	13.8	15.4	16.9	18.3	19.7	21.3
	PACU	8.0	9.0	10.0	11.0	12.0	13.0	14.0	15.0	11.1	11.7	12.0	12.0	12.6	12.8	13.3	13.7
	Total Cost	90.0	75.9	**71.1**	74.0	81.2	91.6	103.6	115.8	135.4	109.1	89.1	79.3	**76.6**	80.8	86.9	95.4
	Cost/LB	3.5	4.5	6.8	6.1	4.0	3.1	2.7	2.5	6.4	5.6	4.5	3.8	2.9	2.5	2.2	2.0
Avg HR		10.2	11.9	13.6	15.3	16.9	18.5	19.9	21.3	10.2	11.9	13.6	15.3	16.9	18.5	19.9	21.3
Avg PACU		8.8	9.8	10.8	11.8	12.8	13.8	14.8	15.8	11.9	12.4	12.5	12.4	12.6	12.9	13.2	13.7
Avg Total Cost		128.1	107.9	95.6	**91.2**	92.8	98.8	107.8	118.7	185.7	155.9	130.6	112.6	101.9	**98.4**	100.8	106.0
Avg Cost/LB Ratio		3.8	4.3	5.7	6.9	6.3	4.7	3.6	3.0	6.6	6.4	5.7	5.1	4.5	3.9	3.4	2.8

Notes: Stochastic Patient Sample Path (# simulation instance = 1000) and Deterministic Durations, OR idle cost $C^d = 1.5$, OR overtime cost $C_1^o = 2$, PACU over-staffing cost $C_2^o = 1$, and $6 \leq R_1 = R_2 \leq 13$

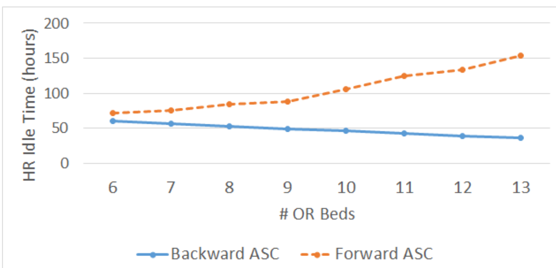
Figure C.7: BackwardASC$_{\mathbf{Alt}}$ vs. ForwardASC



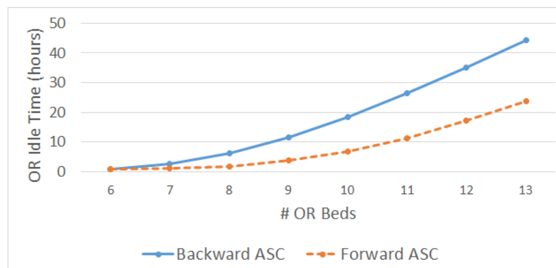
(a) # HRs by # ORs



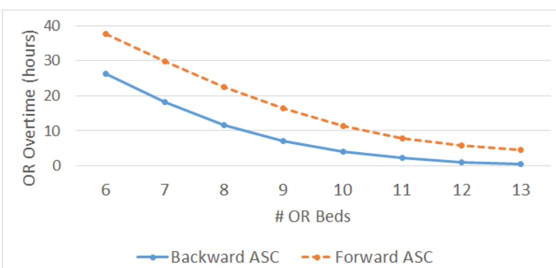
(b) # PACUs by # ORs



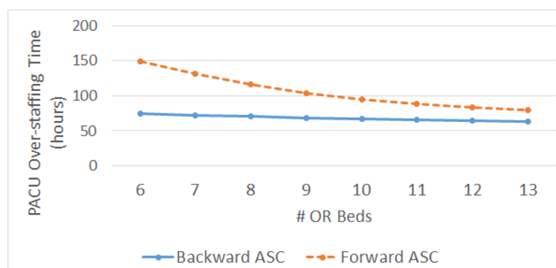
(c) HR Idle Time



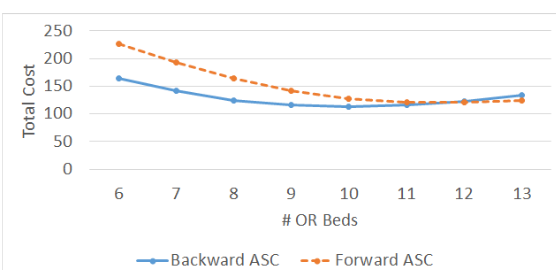
(d) OR Idle Time



(e) OR Overtime



(f) PACU Over-staffing Time



(g) Total Cost

Notes: Results of computational experiment with stochastic Monday patient sample path: BackwardASC$_{\mathbf{Alt}}$ versus ForwardASC (# simulation instance = 1000).

C.9 Extension: Implementation of Surgeon Scheduling on Capacity Planning

Block/rotation scheduling (i.e., scheduling surgeons to different services for each week or month) is an important scheduling problem. Block schedules must satisfy coverage needs of the system along with individual training/administrative requirements to fulfill each surgeon's needs (Lemay et al., 2017). For scheduling purpose, we may also consider surgeon ID since every surgeon has one's own skill sets and experience that affect the duration of the surgical procedure.

In our patient flow data, we are able to observe Surgeon ID for each patient. As listed in Table C.12, each surgeon prefer a certain weekday in general. For example, Surgeon #33 conducted 57 surgeries out of 58 on Monday while Surgeon #32 performed 109 surgeries out of 115 on Tuesday. If the ASC managers have such information in planning capacity, it is beneficial to consider the information as they work as additional constraints in our model that may increase the desired level of capacity in each stage.

Table C.12: Count of Patient Case IDs by Surgeon and by Weekday

Preferred Day	Surgeon ID	Mon	Tue	Wed	Thu	Fri	# Case ID
Mon	33	57		1			58
	91	22		18		12	52
	29	40			2		42
	37	30		1	7		38
	51	30	4	1			35
	67	32					32
	12	22	1		4		27
	70	22			4		26
	11	13	1		2	1	17
	68	10	1				11
	46	5	3			2	10
	26	9					9
Tue	32		109			6	115
	72	1	47	4	10		62
	77		23	9		5	37
	84	2	18	2	4	4	30
	13		27	2			29
	34		26	1			27
	87		16	1		2	19
	76		14				14
	73		6		5	1	12
	69		7				7

Notes: Surgeons who prefer Monday and Tuesday are listed for illustrative purpose.