Impact of Workload and Resource Availability on Hospital Productivity

IMPACT OF WORKLOAD AND RESOURCE AVAILABILITY ON HOSPITAL PRODUCTIVITY

A DISSERTATION PRESENTED
BY
JILLIAN ALEXANDRA BERRY JAEKER
TO
ANITA L. TUCKER
ROBERT HUCKMAN
ANANTH RAMAN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF BUSINESS ADMINISTRATION
IN THE SUBJECT OF
TECHNOLOGY AND OPERATIONS MANAGEMENT

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS
MAY, 2014

Anita L. Tucker Jillian A. Berry Jaeker

IMPACT OF WORKLOAD AND RESOURCE AVAILABILITY ON HOSPITAL PRODUCTIVITY

ABSTRACT

In this collection of essays, I develop a deeper insight into how to incorporate the behavioral aspects of medical care into resource allocation and decision making. Clinicians can, and do, react to variations in the state of the hospital (e.g. an increase in patient load or a decrease in available equipment), by altering patient care in an attempt to meet the needs of all patients. Understanding how these behavioral responses to changing hospital conditions affect care and resource use is necessary for the efficient allocation of equipment, workers, and other resources so as to maximize patient throughput and quality of care.

Contents

1. Operations Management in Healthcare **1**

1.1 INTRODUCTION 1

1.2 THEORETICAL AND PRACTICAL SIGNIFICANCE 2

1.3 OVERVIEW OF DISSERTATION RESEARCH 3

2. Hurry Up and Wait: An Empirical Study of the Spillover Effects of Workload on Patient Length of Stay **5**

2.1 INTRODUCTION 5

2.2 CLINICAL PATIENT CARE PROCESS 9

2.3 RELATED LITERATURE 10

2.4 HYPOTHESES 12

2.5 DATA 16

2.6 ECONOMETRIC SPECIFICATION 19

2.7 RESULTS 24

2.8 DISCUSSION AND CONCLUSION 36

4. Priority and Predictability **73**

4.1 INTRODUCTION 73

4.2 CLINICAL PATIENT CARE PROCESS 75

4.3 RELATED LITERATURE 77

4.4 HYPOTHESES 79

4.5 DATA 82

4.6 ECONOMETRIC SPECIFICATION 86

4.7 RESULTS 88

4.8 DISCUSSION AND CONCLUSION 94

3. Increased Speed Equals Increased Wait: The Impact of a Reduction in Emergency Department Ultrasound Order Processing Time **42**

3.1 INTRODUCTION 42

3.2 LITERATURE ON THE IMPACT OF WORKER DISCRETION ON OPERATIONAL PERFORMANCE 45

3.3 STUDY SETTING 47

3.4 HYPOTHESES 49

3.5 DATA AND ECONOMETRIC SPECIFICATION 53

3.6 RESULTS 61

3.7 ROBUSTNESS 65

3.8 DISCUSSION 67

3.9 CONCLUSION 71

References **97**

Appendix **105**

Author List

Chapter 2 was coauthored with Anita L. Tucker

Chapter 4 was coauthored with Anita L. Tucker and Michael H. Lee

Listing of Figures

2.5.1	Number of visits & LOS of 15 most common medical/surgical DRGs (LOS>=3 days)	19
2.6.1	Summary statistics	21
2.6.2	Correlations among medical and surgical occupancy levels	23
2.7.1	Complementary log-log regression of the effect of inpatient occupancy on Day of Admission on the hazard at time (t)	25
2.7.2	OLS regression of the effect of inpatient occupancy on Day of Admission on ln(LOS)	27
2.7.3	Complementary log-log regression of the effect of inpatient occupancy on Day of Discharge on the hazard at time (t)	29
2.7.4	OLS regression of the effect of inpatient occupancy on Day of Discharge on ln(LOS)	31
2.7.5	Average impact of a 5% change in Day of Discharge occupancy on average LOS in hours	32
2.7.6	Discharge occupancy on medical patient LOS	33
2.7.7	Discharge occupancy on surgical patient LOS	33
4.2.1	Patient care process through hospital	77
4.5.1	Summary statistics	85
4.5.2	DRGs used and number of patients of each type	85
4.7.1	Poisson regression of number of SS admissions (census levels)	89
4.7.2	Poisson regression of number of SS admissions (occupancy levels)	90
4.7.3	Poisson regression of number of surgical discharges (census levels)	91
4.7.4	Poisson regression of number of surgical discharges (occupancy levels)	92
4.7.5	Probit model of the Probability that the day of surgery is after the day of Admission	83

3.3.1	Timeline of changes in U/S ordering process at Flagship and Community.	49
3.5.1	Summary statistics	54
3.6.1	Impact of process change on probability of U/S	62
3.6.2	Impact of process change on waiting times	63
3.6.3	Impact of process change on clinical quality measures	64
3.6.4	Impact of process change on use of other medical tests	65
3.7.1	Impact of process change on U/S across years	66
3.7.2	Impact of process change on U/S use across complaints	67
A.1	Buildup of complementary log-log models for septicemia/severe sepsis w/o MV 96+ hours w/MCC (DRG 871)	106
A.2	Confirmation that patient characteristics did not change pre-/post-change	108
A.3	Categories and medical indications, as described by the American College of Emergency Physicians, for which and emergency U/S is warranted	109
A.4	Symptoms associated with U/S use	110

FOR CHRISTOPH, MOM, AND DAD.
YOU ARE MY PAST, PRESENT, AND FUTURE.
ALL OF THIS IS BECAUSE OF YOU.

Acknowledgements

OVER the last six years at Harvard Business School, I have been fortunate to work and interact with some truly amazing and gifted people. These are the people who have helped me grow as a researcher and enriched my life. Whether through assistance on econometrics, or just chatting over tea, I am thankful to all of these people for making this time so special and fulfilling.

Anita Tucker. I could not have hoped for a better advisor. From the first day I met you, you have combined the perfect combination of challenging me to be my best, while always being supportive. You gave me the freedom to make mistakes, knowing that what I learned from them would be more valuable than anything else you could teach me. The hours spent in your office were some of my most valuable over the last six years. I am so honored to have been your first advisee, and I hope I can pay forward all that you have done for me.

Ananth Raman, you have helped me to keep remembering the big picture through all the details, in both research and life. You have taught me that being a good researcher is about more than just spending all of my time in front of data. But mostly, it is your analogies between research and life that I remember most and am most thankful for. You were right – I had to date a lot of ideas before I married the right one (it worked for the man as well).

Rob Huckman, thank you for always being there when I needed one more set of eyes to read a paper, or when I was trying to discover a new data source. Your ideas and support always gave me the confidence to keep moving forward. And your optimism let me know it would all work out in the end.

Ryan Buell, you were equivalent to a fourth advisor. Your generous kindness and help are truly appreciated. That you want us all to succeed is clear, and I am lucky to have benefited from it.

Bill Simpson and Sarah Woolverton, both of you are the reasons my analyses worked. Without you, I would have no results.

To James Orlin and Micheal Metzger, thank you for introducing me to the joys of operations management.

Lee Fleming, thank you for being on my side during a big transition. I will never forget your call assuring me that I was making the right decision.

To the entire TOM faculty, you are all amazing. How I will miss seeing everyone at those Thursday seminars. And to all of my professors at MIT and Harvard for teaching me all that I know, and helping me grow into the person I am now.

Jen Mucciarone, thank you for always being there for me. You have supported me in more ways than I can ever express. Without you, this would not have been possible. I will miss popping into your office to catch up on life. And to the rest of the doctoral office (past and present), I am so grateful for all of your support (and the many First Friday's) throughout this journey.

Ethi Al-Mahdi, you are the faculty assistant who made everything run smoothly, and for a doctoral candidate, that is something I can never thank you for enough.

I appreciate all of the helpful feedback and suggestions I have received from the seminar participants at the Wharton School, Boston University, Kellogg School of Management, Yale University, University College London, and Georgetown University. And to all of the conference participants throughout the years at INFORMS, MSOM, Decision Sciences, Industry Studies Association, and POMS.

To Rebecca (Becca) Oman, Emily Iacobucci, and Minh Huynh-Le, thank you for being such amazing friends, always having a free couch, and reminding me there is life outside my doctoral program.

Throughout my tenure in the doctoral program, I have had the fortune of working and interacting with some wonderful doctoral student colleagues (and partners). Whether bouncing research ideas off each other or laughing over dinner, it was nice to know we were all in this together. In particular, thank you to Sen Chai, Fern Jira, Nate (and Emily) Craig, Luciana Silvestri, Chris Small (and Saundra Small), Sujin Jang, Matt Lee, Clarence Lee, Joseph Ahn, Pat Satterstrom, Brad Staats, Sam Barrows, Frank Nagle, Anil Doshi, Maria Ibanez, Bill (and Melissa) Schmidt, Hummy Song, Song-Hee (Hailey)

Kim, Tom Best, Hessam Bavafa, Zeshawn Beg, Tiona Zuzul, Abigail Allen, and Lisa Kwan.

Christoph Jäker (aka Jaeker/Jaker), you are the best thing I got out of HBS. I am so glad I talked to you in that hallway. Thank you for all of your support and loving the "smart girl." I cannot wait to see where life brings us next.

To my brother James, thank you for keeping me grounded and being proud of me – it is all a little sister ever wants.

Finally, I have to thank my parents, James and Cynthia Berry. You always knew the doctorate was the right thing for me, but you let me figure it out on my own. You have guided and supported me, and shown me what unconditional love truly is. I love you both so much. And Dad, I may be a doctor now, but I promise that I will still always laugh at your jokes.

1

Operations Management in Healthcare

1.1 Introduction

As healthcare costs in the US and other developed countries continue to rise, there has been an emphasis on improving the efficiency of healthcare delivery by reducing costs while increasing quality of care. In doing so, the goal is to be able to provide better care to more people. Although current operations management theories and improvement strategies can guide these changes, healthcare delivery settings and other complex service environments have unique characteristics which have not been incorporated into earlier operations management work. Specifically, healthcare is delivered by people who have high levels of discretion over the care a patient receives, both in terms of quantity and quality. Yet, the factors that affect the quantity and quality of care that a patient receives are little understood despite being crucial for improvement strategies.

In this collection of essays, I hope to develop a deeper insight into how to incorporate the behavioral aspects of medical care into resource allocation and decision making. Clinicians can, and do, react to variations in the state of the hospital (e.g. an increase in patient load or a decrease in available equipment), by altering patient care in

an attempt to meet the needs of all patients. Understanding how these behavioral responses to changing hospital conditions affect care and resource use is necessary for the efficient allocation of equipment, workers, and other resources so as to maximize patient throughput and quality of care.

1.2 THEORETICAL AND PRACTICAL SIGNIFICANCE

Traditional operations models have generally made the assumptions that the time to complete the work on a particular "unit of work" is a random variable with a constant mean, and that each unit of work is independent from each other unit (Dallery and Gershwin 1992), and these assumptions have been used in healthcare research (Green and Nguyen 2001). However, more recent research suggests that the time required to complete a unit of work is not constant, but is instead influenced by the workload (KC and Terwiesch 2009, Powell and Schultz 2004, Schultz et al. 1999), and availability of resources (Hopp et al. 2007, KC and Terwiesch 2012), though the direction of the change and quality of the work (Kuntz et al. 2013, Oliva and Sterman 2001) may not be constant.

Analytically, it has been shown to be optimal to provide lower service time/quality as demand, and consequently, queuing, increases in a system (George and Harrison 2001, Ha 1998, Stidham and Weber 1989). In healthcare, this behavior has been empirically observed in some settings (Batt et al. 2012, Chan et al. 2012, KC and Terwiesch 2012, Kim et al. 2012) (Add Diwas paper), but it is not always consistent when there are periods with extended or very high levels of demand (KC and Terwiesch 2009, Kuntz et al. 2013). Instead, in these situations, service time actually increases. Moreover, when service quality decreases, due to the nature of healthcare, patients can end up being readmitted, resulting in additional processing time than the time saved (KC and Terwiesch 2012). Each type of change in care could have implications for other patients in the hospital system, and I attempt to provide insight into when each behavioral change (increasing or decreasing service and processing times) occurs, as well as the resulting effect on the primary and other patients.

In addition to understanding the effects of the behavioral response to workload and availability of resources, it is also important to identify any other moderating factors. In healthcare, there are a variety of patients that receive care, and these patients can have markedly different characteristics and needs (Eddy 1984). As a result, the demands of each patient could affect how clinicians alter treatment based on environmental factors. For example, KC and Terwiesch (2012) found that early discharge of less severe patients

under high load may be optimal, and Kim et al. (2012) identify patient types that can be admitted to lower intensity inpatient services when the intensive care unit becomes busier. However, this is an area where much still remains to be explored, and I have tried to identify additional factors that could influence responses to the hospital environment.

1.3 OVERVIEW OF DISSERTATION RESEARCH

In three chapters, my dissertation empirically analyzes how availability and demand for resources affects use and patient care, and how these effects impact other patients within the hospital.

In Chapter 2, entitled "Hurry Up and Wait," we use patient data from 203 California hospitals to explore two different effects of high occupancy – congestion and workload smoothing – on patient length of stay (LOS). We differentiate the conditions under which each effect dominates within a unit, and if these effects spillover to other inpatient units. We find that high occupancy at the beginning of a patient's stay is associated with a longer LOS, consistent with congestion effects, while near the end of the stay workload smoothing effects dominate, resulting in a decrease in LOS. We also show that these effects spillover across hospital units, i.e.: high medical patient occupancy at the beginning of a surgical patient's stay results in the increased LOS associated with congestion effects. Our findings show that high hospital occupancy effects vary depending on the stage of the patient's stay.

In addition to showing that workload related effects differ across a patient's stay, by examining different patient types (i.e., medical, surgical, and obstetrics), we can provide additional explanations for the mechanism of action. Specifically, delays in treatment due to queuing and/or mental fatigue accounts for about half of the increase in LOS due to high early occupancy. The other half of the delay is the result of a prioritization of discharging patients nearing the end of their stays, and is a purely behavioral response to high workload. Overall, this work emphasizes the importance of studying the hospital as a system when allocating resources and setting target occupancy levels: optimal behavior and design for all inpatient units may not be optimal for the hospital as a whole.

Chapter 3, "Priority and Predictability," explores how patient admission characteristics moderate the effects of high workload and demand. Specifically, we study the response to incoming patients that are scheduled versus those that are emergent, by measuring the probability of admission for emergency and scheduled surgical patients. Furthermore, we analyze the effect of incoming patients, by type, on the discharges of

currently convalescing inpatient surgical patients, as well as any delays in the start of care for the incoming patients, depending on admission type. The results of this study show that while scheduled surgical patients, who are less emergent, are more likely to be postponed as the occupancy nears the hospital maximum, are also more likely to cause the early discharge of currently convalescing patients as occupancy increases. In addition, high numbers of scheduled surgical admissions increases the probability that an incoming emergency patient's care will be delayed. The results suggest that scheduled surgical patients are used as a lever to manage hospital capacity. However, the patients that should be receiving prioritization are being delayed as a result, which we theorize is due to the lack of predictability. Therefore, increasing the predictability of emergency admissions, which is possible by using early information, could allow for more efficient hospital capacity management and better quality of care.

Finally, chapter 4, "Increased Speed Equals Increased Wait," explores another factor that can affect worker behavior: resource availability. Often, a suggestion to increase hospital throughput and alleviate the above workload related effects, is to allocate resources that reduce average processing time, such as an additional bed in response to occupancy effects on LOS. However, theory suggests that in settings with high work discretion, reduced processing times can, paradoxically, increase congestion (decrease throughput) due to an increase in service levels. This paper empirically tests this theory by employing a natural experiment in two emergency departments (EDs) staffed by the same physician group, where one ED, but not the other, reduced the processing time for ordering an ultrasound (U/S). We show that reducing the time to complete an U/S, a medical test, does increase the probability of an U/S being ordered. The increased U/S usage increased the LOS in the ED of patients who now receive these additional U/S, as well as patients who receive other radiological tests due to the shared radiology services. Furthermore, the process change resulted in an increase in average waiting time to enter the ED for all patients. We do not observe the expected increase in ED quality of care, as measured by the number of hospital admissions and readmissions to the ED. The results show that increased resource availability may not provide better care, and could actually result in an even greater use of resources. These results provide an explanation for why healthcare costs continue to rise, and the importance of appropriate resource allocation.

J. Berry Jaeker, Tucker, A.L. 2014. Hurry Up and Wait: An Empirical Study of the Spillover Effects of Workload on Patient Length of Stay. Working Paper. Harvard Business School. Boston, MA.

2

Hurry Up and Wait: An Empirical Study of the Spillover Effects of Workload on Patient Length of Stay

2.1 INTRODUCTION

HEALTHCARE spending in the US has increased from $256 billion (9.2% of GDP) in 1980 to $2.7 trillion in 2011 (17.9%), with hospital care accounting for nearly a third of these expenditures (Centers for Medicare & Medicaid Services Office of the Actuary National Health Statistics Group 2012). Consequently, there has been a push to make hospitals more efficient (Sebelius 2013). Given their high fixed costs (Roberts et al. 1999), many hospitals have responded to this cost pressure by increasing occupancy levels. For example, between 1997 and 2006, the average acute care medical/surgical occupancy in California hospitals increased by more than 35% due to an increase in patient demand concomitant with a reduction in the number of medical/surgical beds (State of California OSHPD Health Information Decision 2008).

Although increasing resource utilization through higher occupancy might seem like a straightforward way to improve efficiency, it places a high workload on staff and

resources. The operations literature yields conflicting results regarding the impact of high workload in hospitals, and other service settings. High workload can increase (Kuntz et al. 2013) or decrease a patient's length of stay (LOS) (Anderson et al. 2011, Chan et al. 2012, KC and Terwiesch 2012). Increases stem from queuing effects due to the variability of demand on the system and employees, while decreases stem from the typical human response of smoothing one's workload when faced with high demand. Thus, it is unclear whether high workload will cause an increase or decrease in LOS, and what factors determine this relationship. Queuing-related increases in LOS may dominate workload smoothing-related decreases in LOS under certain conditions while the reverse may be true under others. Another limitation in the current understanding of the relationship between workload and LOS is that prior studies have primarily focused on the effect of an inpatient unit's occupancy (generally the ICU) on the time patients stay within that unit (e.g., Chan et al. 2012, KC and Terwiesch 2012), and not on the entire LOS in the hospital or how the workload-related effects of other inpatient units "spillover." However, the interconnected nature of hospital units (Gittell et al. 2000), makes it likely that workload-related effects can spillover across different types of hospital units such that occupancy of one type of patient impacts the LOS of other types of patients. In this study, we identify the conditions in a hospital under which each workload-related effect dominates, the degree to which they spillover across patient types, and their implications on LOS and throughput.

More research is needed to shed insight on the complex relationship between employee response to workload and its impact on throughput times and resource use (Boudreau et al. 2003, Hopp et al. 2009, Hopp et al. 2007, Schultz et al. 2003), particularly in the healthcare setting (Anderson et al. 2011, KC and Terwiesch 2009, KC and Terwiesch 2012). Improved understanding of the different types of effects of inpatient-unit level and hospital level workload is crucial to increasing hospital efficiency. For example, failing to properly account for the impact of patient load on LOS can result in a non-optimal allocation of labor (Green et al. 2011) and physical resources (Green and Nguyen 2001, Shapiro 1996), and result in poorer performance (Anderson et al. 2012, Debo et al. 2008, Kuntz et al. 2013, Tan and Netessine 2014).

Using patient level data for 182,574 patient visits from 203 hospitals in California, we refine and characterize workload's different effects on patient LOS. We first differentiate workload-related effects into two broad categories: "congestion" and "workload smoothing." Congestion refers to a slowdown effect of high workload due to queuing for

shared resources and mental strain on employees, while workload smoothing describes a "speeding up" behavioral response of shortening service times and prioritizing the discharge of patients near the ends of their stays. Second, we quantify the direct and spillover effects of congestion and workload smoothing on LOS. We categorize patients into one of three clinical classifications (medical, surgical and obstetrics, which we refer to as type) and calculate occupancies for each type. We then compare how of the type and timing (i.e., beginning or end of stay) of occupancy affects the total LOS of patients with one of five medical diagnoses or five surgical diagnoses. By analyzing both the direct and spillover effects related to "own" and "other" patient type occupancy, respectively, we are able to disentangle and compare the changes in LOS related to congestion and workload smoothing. In addition, we exploit that medical physicians provide some care for surgical patients, but the reverse does not hold, to measure how the congestion and workload smoothing effects are moderated by physician discretion. Finally, we supplement our quantitative analysis with interviews of nurses and physicians from California hospitals to better understand the mechanisms through which workload impacts LOS.

We find that for newly admitted patients, a 5% increase in occupancy of their own type is associated with a 0.9% increase in LOS on average, consistent with a congestion effect. Additionally, these effects spillover from one type of patient to another, particularly when physicians are shared (i.e., a high volume of surgical patients increases LOS for newly admitted medical patients). However, the increase in LOS was smaller (0.5% increase in LOS for each 5% increase in occupancy) than for own type occupancy.

An increase in own type occupancy on the day of discharge is associated with a U-shaped effect on a patient's LOS. These results are indicative of a workload smoothing effect where clinicians prioritize discharging patients who are approaching the end of their hospital stay to reduce workload and make room for incoming admissions. However, once workload becomes very high, congestion effects become more pronounced and LOS increases as staff lack the spare capacity necessary to undertake discharges, which can be time-consuming. We also find that this prioritization of discharges impacts newly admitted patients since it can result in delays in the start of their treatments. The LOS of newly admitted medical patients is on average 0.6% longer for a 5% increase in own type occupancy compared to a 5% increase in surgical occupancy due to the prioritization of dischargeable patients. Our results show that workload smoothing not only affects patients near the end of stay, but also amplifies the

congestion effects for patients at the start of their stay. Thus, there is a tension between the managerial need to reduce workload—which focuses attention on the patients near the end of their LOS—and the clinical need to care for the newest (and thus sickest) patients in the inpatient unit.

We find a similar relationship between LOS and occupancy from the other type of patients, even if the patients do not share physicians or nursing staff, though the effect is smaller than for own type patients. These results suggest that some of the behavioral response of workload smoothing spills over across patient types. Observational data and interviews with practitioners confirm that there is pressure on clinicians to reduce occupancy levels as the hospital becomes busier, regardless of patient type.

Our work makes several contributions to the growing operations management literature on the impact of employee behaviors and workload on organizational performance. First, we analyze the effects of occupancy at the beginning and end of patients' hospital stays, and find that different effects dominate at different times, a concept that we believe has not been fully explored in previous work. Second, we identify and quantify spillover effects of workload across different types of patients on LOS. We find that workload-related effects spillover across patient types, which suggests that future studies should include possible spillover effects if they are to fully account for workload's impact. Relatedly, we show that there is a workload smoothing spillover effect, which is a surprising behavioral response not theorized to occur when other type workload is high and resources are shared. Our study underscores the importance of allocating resources across the hospital to achieve optimal hospital-level, rather than department-level performance. Third, by quantifying the different workload effect types and how they spillover, our work enable us to separate out the effects related to unit- and hospital-level demand from those related to clinician behavior, including the effect of prioritizing discharges over admissions. In the discussion section, we provide an example of how these results can impact the capacity and discharge decisions of a hospital. The disentanglement of congestion and workload smoothing effects has significant implications for operational solutions for reducing hospital LOS: congestion effects, the focus of most previous work, are primarily addressed through capacity decisions, while the behavioral aspects of workload smoothing are better confronted with the standardization of work (e.g. with checklists) and careful allocation of responsibilities.

2.2 CLINICAL PATIENT CARE PROCESS

To understand why and how we categorize patients by type, it is critical to understand the care process of a patient through the hospital. Based on a patient's diagnosis on arrival, the patient is assigned to a specific hospital service, which is the specialty responsible for caring for the patient. Three broad categories of patient services exist: medical, surgical, and obstetrics. Although some hospitals, particularly larger, academic centers, have subspecialties, these services are not standardized across hospitals, so we use the above broad clinical service categorizations. Each service has its own physicians, and the primary decisions regarding care are generally segregated by service (Gittell et al. 2000, Meyer 2011). The segregation is also heightened by the fact that the different types of physician services typically have separate financial silos(Glouberman and Mintzberg 2001). Ancillary services, such as radiology, pharmacy, laboratory, and transport, are shared among all patients in the hospital, while beds, nursing, and physicians overlap to differing extents based on the service and hospital. For example, while obstetrics patients are usually completely separated from the general population, medical and surgical patients have more overlap in care. In fact, it is common for lower acuity medical and surgical patients to be co-located on a general nursing unit. Furthermore, although the ultimate care decisions for surgical patients are made by surgeons, medical physicians are often involved in following those patients as well. However, it should be noted that the reverse does not occur, i.e. a surgeon is rarely involved in the care of a medical patient. We explain in section 4 how we use these differences in shared resources to distinguish between congestion and workload smoothing effects.

We use the occupancy of each patient type at the start and end of a patient's stay as measures of workload. We use these two time points because our discussions with physicians and nurses have shown that they are the periods when the majority of treatment decisions are made and resources are used, and thus when occupancy is most likely to affect LOS. Therefore, we refer to the *occupancy* of medical (surgical, obstetric) patients in the hospital on the day of admission (discharge) as medical (surgical, obstetric) occupancy at admission (discharge). We will also control for the workload from incoming patients. We refer to the occupancy of medical (surgical, obstetric) patients awaiting admission as the medical (surgical, obstetric) *admission rate* on the days of admission and discharge. Please note that to account for differences in hospital sizes, we use occupancy and admissions rate—which are percentages of a maximum--rather than the cardinal number of current and admitted patients in the hospital.

To categorize each patient into the above types, we rely on the Center for Medicare and Medicaid Service's (CMS) diagnostic related group (DRG) and major diagnostic category (MDC) classifications. Upon completion of a patient's hospital admission, each patient is assigned one of 746 DRGs based on the clinical diagnosis and primary needs of the patient, including the patient's severity level (Office of Inspector General 2001). While there is some variation among patients with the same DRG, CMS treats it as minimal and uses DRGs to calculate payments for patient care irrespective of the costs incurred (Office of Inspector General 2001). Each DRG is a patient subtype within one of 25 MDCs, such as pregnancy, childbirth, and puerperium, which we classify as obstetrics. In addition to the MDC, each DRG is broadly defined to be of type medical or surgical. Of the remaining non-obstetrics patients, we classify each as surgical or medical based on his or her DRG.

2.3 Related Literature

Our research focuses on understanding how the workload in a hospital affects patient LOS, or service times. Traditional operations models have generally made the assumptions that the time to complete the work on a particular "unit of work" is a random variable with a constant mean, and that each unit of work is independent from each other unit (Dallery and Gershwin 1992). These assumptions have been extended to research in patient care settings. However, recent research suggests that the time required to complete a unit of work is not constant, but is instead influenced by the overall workload, though the direction of the change may not be constant. Instead, this stream of research describes two competing factors at work: *congestion* due to queuing for resources under high workload and *workload smoothing*, which is workers' behavioral response to reduce their workload when faced with high demand periods.

Queuing theory states that when there is variation in service and arrival times, both of which occur in hospitals (Eddy 1984), queuing will occur. As utilization approaches 100%, queue length and overall throughput time increase dramatically. In healthcare settings, studies have shown that as patient load becomes higher, LOS can increase (Green and Nguyen 2001) and quality of care decrease due to the resulting delays in treatment (Chalfin et al. 2007) and mental strain (Kuntz et al. 2013). In addition, increased load in a healthcare setting has been shown to increase the number of interruptions, which often results in repeat set-up periods and re-work due to errors (Tucker and Spear 2006). Prolonged heavy load in healthcare has also been linked to

worker exhaustion that results in slowdown and a propensity for errors (KC and Terwiesch 2009). As a result, slow-down and longer LOSs may occur when workload is heavy. We collectively refer to these slow-down effects as congestion.

Other research has found that the relationship between capacity utilization and throughput time predicted by queuing theory can be altered when employees have discretion over how they complete their tasks (Debo et al. 2008, Hopp et al. 2009, Hopp et al. 2007, Jouini et al. 2008). Specifically, workers can speed up or slow down if they have discretion over two key behaviors: how many tasks they perform for customers (Batt et al. 2012, Hopp et al. 2007, Oliva and Sterman 2001) and how long they take to perform those tasks (KC and Terwiesch 2009, KC and Terwiesch 2014, Schultz et al. 1998). Thus, workers have levers to increase or decrease processing times for customers, which in turn impacts the link between workload and overall throughput time.

Analytical studies provide support for the notion that workers have reason to reduce customer processing times under high workload. When customers are queued up waiting for service, the average utility of all of the customers can be increased by decreasing the time spent per customer (George and Harrison 2001, Ha 1998, Stidham and Weber 1989). This is because the cost of increased effort and/or reduced service level is offset by the reduction in waiting times (George and Harrison 2001, Stidham and Weber 1989). Therefore, it is optimal for the system for servers to reduce processing times for each individual customer (Ha 1998), and to only provide more services—which is what customers would prefer—when there is enough server capacity (Hopp et al. 2007).

Empirical studies of employees' response to workload have identified some of the specific mechanisms by which workers reduce processing times. Employees whose number of tasks is fixed, and whose work is standardized, have been shown to speed up when faced with rising inventory levels (Schultz et al. 1999, Schultz et al. 1998). In settings where employees have the discretion to adjust the number of tasks or depth of work that they perform for a customer, research has found that as workload increases employees will first work faster (Kuntz et al. 2013, Tan and Netessine 2014). If queuing continues, they will then "cut corners" by omitting certain tasks to reduce waiting time for queuing customers (Oliva and Sterman 2001). We refer to this behavior as workload smoothing. In a healthcare setting, high patient load at the end of a patient's stay has been shown to result in worker speed up for that patient (KC and Terwiesch 2009). Under periods of extended high workload, staff cut corners by prematurely discharging

patients from the ICU (KC and Terwiesch 2012, Kim et al. 2012) or hospital (Anderson et al. 2011), or stinting on documentation (Powell et al. 2012). The corner cutting can negatively impact the quality of work, as evidenced by higher mortality rates (Kuntz et al. 2013), lower reimbursement (Powell et al. 2012), and higher readmission rates (Anderson et al. 2012) when inpatient occupancy is high.

Congestion effects work in the opposite direction of workload smoothing behavior. If the rate of incoming demand is greater than the rate at which the system can process that demand, the queue will lengthen and average throughput time will increase (Little 1961). Thus, even though a worker may speed up to reduce a particular patient's LOS, if the change in service rate is not as high as the change in workload, then on average, high workload will result in an increase in LOS. However, to our knowledge, there is minimal research which distinguishes when slow down due to congestion or speed up due to workload smoothing will occur in a hospital setting.

To reconcile this tension in the direction of change in LOS due to increased workload, we examine whether workload has a different effect on LOS at different times during a patient's stay, and as a result of high workload from different patient types. Previous research has primarily focused on workload at the end of a patient's stay within an inpatient unit (generally in the ICU) (KC and Terwiesch 2012, Kim et al. 2012), and has not fully explored the effects of workload at the start of the stay. Through our work, we are able to show that high levels of hospital occupancy do not have a consistent impact on patients. Our results highlight the importance to both theory and practice of teasing apart the "average" effect of workload to determine its specific impact on individual patient types.

2.4 HYPOTHESES

2.4.1 OCCUPANCY AT TIME OF ADMISSION

The start of a patient's stay is focused on diagnosis and treatment, which generally requires more staff and physical resources (Clarke 1996) than the remainder of the stay, which is focused on the patient's recovery. During the initial period, there is a required set of time consuming tasks that must be completed, many of which require resources shared by many patients. Based on our discussions and observations of physicians and nurses in California hospitals (described in more detail in the methods section), these include reviewing the patient's history, orders, and allergies; getting vital signs;

observing and documenting any pre-existing skin ulcers on the patient; writing new orders; and beginning treatment. The physicians we observed took between 90 minutes and two hours to write the initial patient orders for a standard patient. Completing admission tasks required 30 to 60- minutes of nursing time.

An implication of the lack of discretion in the set of tasks required by patients at the beginning of their stay is that clinicians are less able to use workload smoothing to omit tasks for their newly admitted patients. Furthermore, as we describe in section 3, under periods of high occupancy, it is likely that work will be delayed by queuing, interruptions, and fatigue and error caused by the mental strain on staff (KC and Terwiesch 2012, Kuntz et al. 2013, Tucker and Spear 2006). Delaying the start of care for critical patients has been shown to increase the overall LOS for a patient, beyond the direct delay in care (Chalfin et al. 2007), likely due to the inability to perform key tasks during the "golden" first hours of a patient's stay (Blow et al. 1999, Buist et al. 2002). The queuing and mental strain associated with high load will affect all care providers and resources within the hospital, so the effects can spillover across patient type. Consequently, as the occupancy of any type increases on the day of admission, we would expect congestion effects to occur and dominate workload smoothing and thus result in a longer LOS. Formally,

HYPOTHESIS 1A (H1A): *Increased occupancy of other type patients on the day of admission is associated with an increased LOS (congestion effect).*

Although newly admitted patients are the most in need of clinical care, given the long time required to complete a new admission, as inpatient workload increases staff may first try to complete important work for their other patients, whose care would otherwise be severely delayed while the clinician completes admission work. This delays the start of the treatment process for newly admitted patients. In particular, clinicians may prioritize discharging patients near the ends of their stays to smooth their workloads. By focusing on getting patients discharged, clinicians can free up physical and mental capacity to care for the newer, sicker patients. In addition, the benefits of discharging a patient near the completion of her care will be apparent immediately, while the effects of any delays in the start of care for a new patient will not come until the patient is ready for discharge days later, so the immediate costs of focusing on newly admitted patients far outweighs the immediate benefits. This highlights the tension between the clinical needs of patients and the managerial and behavioral response to free

up resources. Thus, in addition to the increase in LOS due to congestion, we expect an increase in LOS due to the behavioral response of workers to prioritize care of other patients besides the newly admitted ones (i.e. workload smoothing). However, while congestion affects both same and other patient type, the ability to prioritize discharges over admissions is only present for a patient's primary physician and nurse. Therefore, high occupancy of other patient types is less likely to trigger this behavioral response. Thus, the difference in LOS increase between own and other type workload represents the effect of workload smoothing. Consequently, we would expect increased own type occupancy to result in a larger LOS increase compared to other type occupancy.

HYPOTHESIS 1B (H1B): *Increased occupancy of own type patients on the day of admission is associated with an increase in LOS (workload smoothing and congestion effects) and this increase is greater than the increase in LOS associated with an increase in other type occupancy (workload smoothing effect).*

4.4.2 OCCUPANCY AT TIME OF DISCHARGE

Each patient has an expected LOS associated with his primary condition, as well as patient specific characteristics, such as age, gender, and presence of comorbidities. The decision to discharge a specific patient requires balancing the effects of sending the patient home before he is fully recovered, which increases the chance of readmission (Chan et al. 2012, KC and Terwiesch 2012) versus keeping him longer, which enables a more thorough recovery, but also incurs greater cost for the hospital without additional revenue due to the fixed fee payment based on the patient's DRG. Longer stays also increase patients' risk of hospital acquired infections and medical errors (Hauck and Zhao 2011). We hypothesize that when total occupancy is high (including other type patients) around the time of discharge, the decision to discharge will be affected by more than medical concerns alone.

In some hospitals, the authors have observed that as the hospital's total occupancy becomes high, senior administrators pressure all clinicians to prioritize discharging patients who are ready to go home. If the high occupancy within the hospital comes primarily from other type of patients, the nudge from hospital administrators to discharge one's own patients will have a positive, immediate impact on clinicians as it can result in a discharge, which will reduce a clinician's own workload. For example, discharging a patient eliminates the need to "hand-off" the patient at the end of the shift, a time consuming task that we observed adds to the load of nurses and physicians, and thus provides an additional incentive for early discharges. Even if the administrators do

not actively encourage discharge, clinicians may notice longer queues for shared resources making it more difficult to care for the sickest patients, and thus be incentivized to discharge patients. Clinicians can speed up discharge because they have significant discretion over when discharge-related tasks can be completed. For example, physicians can order a discharge at any time they feel the patient is ready to go home, and nurses can complete many discharge-related tasks (such as patient education) well in advance of when the patient actually leaves the hospital. Finally, some tasks may even be able to be shifted to outpatient follow-up care. Therefore, as the hospital's total occupancy of other type patients increases, we would expect the benefits to clinicians of engaging in workload smoothing to outweigh the costs. As a result, the LOS for a patient near discharge will be reduced through workload smoothing when occupancy of other patient type is high.

However, the ability of a clinician to complete all of the tasks required for an early discharge often requires shared resources. For example, some patients need to be assessed for swallowing ability or stair climbing ability by a therapist before they can be released from the hospital, while others must have a reconciliation of medications before discharge, and this must be performed by the hospital pharmacy. Consequently, at very high occupancy levels of other type patients, patients might have longer waits for discharge tasks to be completed, slowing down the process and leading to longer LOSs. Consequently, we expect that at very high levels of other type patient occupancy, congestion effects will counteract the workload smoothing effects, and the LOS for a patient near discharge will increase. For that reason, we expect a U-shaped effect of other type occupancy on the day before expected discharge on the LOS of a current patient, such that LOS will initially decrease as other type occupancy increases from low levels of workload, but LOS eventually increases at high levels of other type workload.

HYPOTHESIS 2A (H2A): *Increase of occupancy of other type patients on the day of discharge is associated with a U-shaped LOS (workload smoothing to congestion).*

As own type workload increases, a clinician has an even greater incentive to smooth her workload by discharging patients earlier than when other type workload is high because it reduces her own already heavy personal workload (Green et al. 2011, Tan and Netessine 2014), although it may interfere with the start of care of newly admitted patients, as described above. Congestions effects also are more amplified when own type occupancy increases. At very high occupancy levels, clinicians may only be able to

maintain patients (Aiken et al. 2002, Kuntz et al. 2013) and may be too strained to speed up to smooth their own workload, and instead will be more likely to slow down due to fatigue (KC and Terwiesch 2009, Tan and Netessine 2014). Workload smoothing can only occur if there is additional capacity to handle discharges. By definition, as own type occupancy increases, the clinicians will be spread across more patients, and thus will be slowed down by queuing and mental strain. Thus, we expect the following:

HYPOTHESIS 2B (H2B): *Increase of occupancy of own type patients on the day of discharge is associated with a stronger U-shaped LOS than an increase in other type patient occupancy (workload smoothing to congestion).*

2.5 DATA

Our data consists of patient-level records for all inpatient discharges in the state of California from December 2007 through December 2009 (Office of Statewide Health Planning and Development). We sum all patient admissions and subtract all patient discharges for patients admitted in December, 2007 to determine the baseline number of patients in each hospital on January 1, 2008. We restrict our sample to only include patients admitted after December 31, 2007, and before November 30, 2009 since our data only has patients who were discharged by December 31, 2009, and does not include any patients admitted in December, but discharged on or after January 1, 2010, thus making it impossible for us to get an accurate census during that month. Out of the 448 hospitals in the data set, we limit our data to acute care hospitals with (1) at least 3500 patients over the course of our study (23 months) to ensure enough patients per day to calculate reasonable occupancy levels; (2) an average occupancy of obstetrics patients of at least 40% to guarantee more than intermittent obstetric patients; and (3) a 24 hour ED to ensure there are emergency admissions. This resulted in a sample of 203 hospitals. Each admission is its own record and includes the date admitted, date discharged, demographic information on the patient (e.g. gender, race, age), hospital of care, diagnoses (including comorbidities, which we categorize using the Elixhauser Index (Elixhauser et al. 1998)), major procedures, disposition (which is where the patient was discharged to: home, home health services, etc.), if an admission was scheduled or not, and DRG, among others. Consequently, we know whether the patient was medical, surgical, or obstetric, and if the visit was scheduled or emergent. Lastly, we have dates for all major procedures performed, and the LOS of each patient in days. For medical patients, our primary measure of LOS is the day of discharge minus day of admission,

while for surgical patients, we measure LOS as the date of discharge minus the date of surgery. We use these different measures because they are representative of the amount of care each respective patient receives. In particular, surgery may be delayed due to a patient's arrival time (e.g., at night) or another reason that is independent of workload. As a robustness check, we measure LOS for surgical patients as day of discharge minus day of admission.

The hospitals in our data vary greatly in the number of admissions and number of beds, so the patient censuses and number of admissions can similarly vary. To be able to analyze the effects of patient workload across these heterogeneous hospitals, we converted the absolute number of medical, surgical, and obstetrics patients currently in the hospital, as well as the number of incoming medical, surgical, and obstetrics patients, for each date, into rates that are a percentage of the maximum for each patient workload category for that hospital. While we would prefer to calculate occupancy and admission rates based on staffed beds available for a given day, the recorded numbers of staffed beds are at the year level, and often quite different from reality. Consequently, in a manner similar to that employed by Kuntz et al. (2013) we divide the daily census of the hospital (patients admitted that day) by the maximum daily census (daily admissions), and refer to this value as occupancy (admission rate). Specifically, we calculate the 99th percentile of the number of patients receiving treatment in a hospital (or admitted) in a day, by type, over the previous 90 days. We use the 99th percentile to account for any extraordinary circumstances that do not actually reflect the number of beds generally available. For example, to calculate the occupancy for hospital "H" on Monday, June 9, 2008, we first find the 99th percentile of the number of patients who were inpatient in hospital H from March 11, 2008 through June 9, 2008, and label that the maximum. We then divide the number of patients actually in the hospital on June 9 by that maximum and multiply by 100 so that it is a percentage between 0 and 100. Since occupancy is a calculated value, but is crucial to our study, we want to ensure that our results are not sensitive to our definition of occupancy. Therefore, for robustness checks, we calculate two other measures of occupancy. For the first alternate occupancy measure, we calculate occupancy as the maximum of the previous 90 days, but distinguish between weekdays and weekends to account for what is commonly termed the "weekend effect" in hospitals. A second alternate occupancy measure uses the quarter-year instead of 90 days as the range. In both situations, we calculate maximums for weekends and weekdays, so that in the example above the maximum would be the 99th percentile of the

number of patients who were inpatient on a weekday in hospital H in the previous (1) 90 days or (2) quarter 2 of 2008.

We focus our analysis on ten DRGs (five medical and five surgical). We restrict our analysis to these DRGs since they are among the 15 most common medical and surgical DRGs in our sample with an average LOS of at least 3 days and less than 8 days among U.S. Medicare patients (see **Figure 2.5.1**), with a total of 182,574 patient visits. We use the frequency and at least 3-day LOS criteria for inclusion to ensure that we have enough patients and a long enough LOS to detect day-level changes in LOS. To reduce noise in our sample, we focus on DRGs with average LOSs shorter than 8 days because the inherent variability in long LOSs makes it difficult to predict the LOS for those patients. Analyzing ten DRGs provides evidence that our results are robust across patient types. In the DRG classification, the same primary complaint can have different DRG numbers to distinguish the severity level. As a result, each of the ten DRGs in our study represents a relatively homogeneous set of patients with the same specific primary reason for visit and severity level.

To further reduce noise from our analysis of the impact of workload on a patients' LOS, from the 10 DRGs that we analyze we exclude patients who die during their stay since it is very difficult to predict the LOS of a patient who ends up dying in the hospital (Kuntz et al. 2013), and the death of a patient should be uncorrelated with the LOS of the remaining patients outside of the effect of occupancy and only 0.50% of the patients in our DRGs of interest died within the hospitals in our sample. We also exclude patients who are transferred to another hospital, as these patients have LOSs that are impacted by factors not directly related to the occupancy of the hospital, as there is little evidence that transfers take place due to hospital occupancy (Kuntz et al. 2013). Transfers also represent less than 2% of the total patients within these ten DRGs. Moreover, we exclude patients who left against medical advice as it is difficult to predict when they will leave, and they are only 1.10% of the sample. Finally, we restrict our analysis to patients who were emergency arrivals to the hospital. Since scheduled surgeries or visits could potentially be cancelled when high load is present, the patients who are not cancelled could be fundamentally different from the cancelled patients. Meanwhile, emergency patients are by definition random, and thus support that workload is not affecting unobservable patient characteristics and severity levels. Note that while we focus on ten DRGs for calculating effects of workload on LOS, our occupancy and admission rates include all DRGs because the current occupancy levels and admissions

in the hospital are comprised of all patient types, including those who died, were transferred, left against medical advice, and had scheduled admissions.

Figure 2.5.1. Number of visits & LOS of the 15 most common medical/surgical DRGs (LOS>=3 days).

DRG	*Name*	*Number of Patient Visits in our sample**	*Average LOS among U.S. Medicare Pts (Days)*	*Average LOS among patients in our sample₸ (Days)*	*Type*
470	Major joint replacement or reattachment of lower extremity w/o MCC	49,882	3.8	5.06	SURG
871	Septicemia/Severe sepsis w/o MV 96+ Hrs w/MCC	36,059	7.3	8.85	MED
603	Cellulitis w/o MCC	20,913	4.6	5.25	MED
194	Simple pneumonia & pleurisy w/CC	19,718	5.1	5.73	MED
291	Heart failure & shock w/MCC	17,112	6.4	7.0	MED
690	Kidney & urinary tract infections w/o MCC	16,235	4.2	4.86	MED
292	Heart failure & shock w/CC	12,466	4.7		MED
641	Nutritional & misc. metabolic disorders w/o MCC	8,564	3.7		MED
312	Syncope & collapse	6,886	3.1		MED
460	Spinal fusion except cervical w/o MCC	5,186	4.0	7.05	SURG
481	Hip & femur procedures except major joint w/CC	4,698	5.7	5.65	SURG
419	Laparoscopic cholecystectomy w/o CDE w/o CC/MCC	4,491	3.0	3.89	SURG
742	Uterine & adnexa proc. for non-malignancy w/CC/MCC	4,173	4.3	5.39	SURG
494	Lower extremity & humer. proc. except hip, foot, femur w/o CC/MCC	4,107	3.2		SURG
418	Laparoscopic cholecystectomy w/o CDE w/CC	3,465	5.3		SURG
	TOTAL	213,955			
	TOTAL for DRGs of Interest	182,574			

The shaded DRGs are the ones that we analyze. Darker shading are medical, lighter shading are surgical.
*Prior to restricting our sample to account for emergency admissions and LOSs of at least 3 days
₸After restricting our sample to emergent patients with LOS>=3 days
CC=Comorbidities and/or Complications; MCC=Major CC; MV=Mechanical Ventilation; CDE=common duct exploration

In addition to patient level data, we also interviewed three medical physicians at two California hospitals about their patient care processes. We followed up these interviews with observations at one of the hospitals, where the first author followed two physicians through admissions and rounding of inpatient medical and surgical patients. In addition, we interviewed five medical/surgical nurses from the same hospital about what they did to prepare for a new admission.

2.6 ECONOMETRIC SPECIFICATION

The LOS of a patient can be thought of as a survival function, with discharge being equivalent to "failure," or exiting the system (Jenkins 2004, KC and Terwiesch 2012). Survival analysis allows us to predict a patient's likelihood of discharge on any given day based on an underlying hazard function scaled by the patient's and hospital's characteristics. Since our LOS data is at the day level, we use a discrete-time survival analysis. In healthcare settings it is common to allow the baseline hazard to vary with time (KC and Terwiesch 2012), and since our data does not fit a known distribution very well, we include a variable for each day to account for this flexible hazard rate.

We follow Jenkins' (2004) approach and use a proportional hazard complementary log-log (cloglog) regression to model the probability of "failure" (i.e., discharge) at any given time (in our case, for each day). Cloglog is an appropriate model since it can be used with discrete and censored survival times, and it is the best estimator when observation times are discrete but the events occur continuously. In our study we have daily census, admission, and discharge data, but patients are admitted and discharged throughout the day. We model the effect of occupancy as a spline function with two defined knots that represent a busy and very busy inpatient unit, respectively. This gives us the following hazard for patient *i* at time *t*:

$$h_i(t) = 1 - \exp[-\gamma_i \exp(h_0(t))] \tag{2.1}$$

Where $h_0(t)$ is the baseline hazard on day *t* and

$$\begin{aligned}\gamma_i = \exp\big(&\beta_0 + \beta_1 * \mathrm{Controls_i} + \beta_2 * \mathrm{Occupancy_{i,0}} + \beta_3 * \mathrm{BusyOcc1_{i,0}} + \beta_4 \\ &* \mathrm{BusyOcc2_{i,0}} + \beta_5 * \mathrm{Occupancy_{i,d}} * \mathrm{I_d} + \beta_6 * \mathrm{BusyOcc1_{i,0}} * \mathrm{I_d} \\ &+ \beta_7 * \mathrm{BusyOcc2_{i,0}} * \mathrm{I_d}\big)\end{aligned} \tag{2.2}$$

in which β_i is a vector of coefficients, *Controls* is a vector of controls, $Occupancy_{i,0}$ is a vector of medical, surgical, and obstetric occupancy levels for patient *i* on her day of admission (day 0), and $BusyOcc1_{i,0}$ and $BusyOcc2_{i,0}$ are the marginal occupancies (for each occupancy type) associated with being busy and very busy, respectively, on patient *i*'s day of admission. $Occupancy_{i,d}$ is a vector of medical, surgical, and obstetric occupancy levels for patient *i* on her day of discharge (day d), and $BusyOcc1_{i,d}$ and $BusyOcc2_{id0}$ are the marginal occupancies (for each occupancy type) associated with being busy and very busy, respectively, on patient *i*'s day of discharge. I_d is a binary variable equal to 1 if day t is greater than or equal to d-1, and 0 otherwise since the day of discharge workload levels cannot affect the likelihood of a patient going home before that day. This method not

only allows us to take into account that discharge occupancy should not affect the probability of leaving before the end of the patient's stay, but also that patients with the same DRG nonetheless can have different LOSs. For example, we could use the average LOS as the discharge occupancy, but some patients stay longer due to their individual comorbidities and response to medical treatment (Gawande 2007). Control variables include patient age in years, race, gender, payment type, day of week on which the patient arrived, year of the admission, disposition (e.g., discharged to home, nursing home, skilled nursing facility, etc.), and admission rates on day of admission and discharge. As with the occupancies above, we interact the admission rates on the day of discharge by the same indicator variable. **Figure 2.6.1** shows the mean values for the control variables as well as for the main variables of interest.

Figure 2.6.1 Summary statistics (N=203 hospitals)

		Mean	SD
Length of Stay *(Days)*	Total (day discharge – day admission)	6.4	4.5
	Care time (day discharge – day procedure)	6.3	4.5
Occupancy (%)	Medical	78.0	11.8
	Surgical	71.1	15.8
	Obstetrics	59.2	20.0
Control Variables (Continuous)			
Admissions (%)	Medical	60.6	17.1
	Surgical	44.6	25.2
	Obstetrics	46.7	22.9
Age *(Years)*		66.8	19.5

Control Variables (Categorical)		% of Visits			% of Visits
Day of Week [7]	Sunday	12.6	**Payment Type [10]**	Medicare	59.7
	Monday	15.81		Medi-Cal	14.3
	Tuesday	15.24		Private Coverage	18.2
	Wednesday	14.54		Workers' Compensation	0.4
	Thursday	14.15		County Indigent	2.3
	Friday	14.82		Other Government	0.5
	Saturday	12.85		Other Indigent	0.5
Year [2]	2008	50.0		Self Pay	3.8
	2009*	50.0		Other Pay	0.3
Race [7]	White	58.3		Unknown	0.01
	Black	8.2	**Disposition [9]**	Home	55.0
	Hispanic	23.2		Acute Care w/in admit hospital	0.02
	Asian/Pacific Islander	7.3		Other Care w/in admit hospital	0.8
	Native American/Eskimo/Aleut	0.2		Skilled Nursing/Intermed Care w/in admit hospital	2.8
	Other	2.3		Skilled Nursing/Intermed Care at another facility	24.3
	Unknown	0.5		Residential Care Facility	1.4
Gender [2]	Male	41.5		Home Health Services	15.1
	Female	58.5		Other	0.6
				Unknown	0.01

Numbers in **[]** are number of categories for the variable; Shaded categories are most common category of that variable

For medical and surgical occupancy, we define "busy" as an occupancy greater than 85% as this is the average occupancy most hospitals aim for, but over which they consider it to be busy. We define "very busy" as occupancy greater than 93% as this is the value that Kuntz et al. (2013) found to be associated with an increase in the mortality rate. As Figure 2.6.1 shows, obstetric occupancy and admission rates are lower in general, so we use lower occupancies corresponding to the same percentiles to define the variables busy and very busy (70% and 85% respectively). To illustrate equation 2.2, for a patient who experienced a medical occupancy of 87% on the day of admission, the value for *Occupancy* would be 87%, *BusyOcc1* would be 2% (i.e. 87% minus 85%), and *BusyOcc2* would be 0. Analysis in this manner allows the effect of occupancy to change at different occupancy levels.

We restrict our analysis of LOS to patients with the ten DRGs of interest, running separate analyses for each DRG given that each will have a different baseline hazard function. We calculate this baseline hazard rate for each patient for each day, $d=1,2,...D$, where D is the 99th percentile of LOS for the DRG, providing the baseline hazard function associated with time. We consider the 99th percentile of LOS to be the maximum expected LOS since we have observed that patients who have abnormally long LOS often have unobservable medical or social conditions unrelated to patient census that is causing the long LOS (e.g., they do not have a safe home environment). To account for this cutoff, we censor patients with a LOS greater than D days, which can be interpreted as having had no event before the end of the observation period.

Our main variables of interest are medical, surgical, and obstetrics occupancy, as well as our busy occupancy variables, on the day of admission and the day of discharge. In addition, we include medical, surgical, and obstetric admissions on the day of admission and day of discharge as additional controls. Given that we only have midnight occupancy levels, we use the occupancy level at the start of the day of admission (the midnight immediately prior to being admitted) and the start of the day of discharge for our inpatient variables. For our admission rate variables, we use the total number of admissions on the day of admission and on the day of discharge.

Note, while we observe the workload measures on the day of admission and day of discharge, we do not control for the level of busyness on the other days of the stay since we do not expect them to have a significant effect on LOS (see Robustness section for analysis that shows this to be true). Furthermore, we are concerned about over-specification due to the high correlation in occupancy of the same type of patients

between days. As shown in **Figure 2.6.2**; the correlation between day "D" and day "D+1" is 0.83 for medical patients, and 0.83 for surgical patients ($p<0.05$ for both).

Figure 2.6.2. Correlations among medical and surgical occupancy levels (n=112,271)

		Medical Occupancy		Surgical Occupancy	
	Day	*D*	*D+1*	*D*	*D+1*
Medical Occupancy	*D*	1	-	-	-
	D+1	0.83*	1	-	-
Surgical Occupancy	*D*	0.19*	0.23*	1	-
	D+1	0.13*	0.17*	0.83*	1

*p<0.05

Given the set of hazard functions, we can also solve for the patient's survival function $S_i(t)$, which gives the chances of surviving past time t, and equals

$$S_i(t) = \exp\left\{\sum_{d=1}^{t} \ln[1 - h_i(d)]\right\}, where \quad S_i(0) = 1 \; \forall \; i \tag{2.3}$$

As with the hazard function, we can calculate these survival functions for each patient for each day, $d=1,2,..,D$, to yield a survival curve, and find the expected survival time of the patient, or LOS, of a patient with a specific set of characteristics, and the relative effects of each of these characteristics, as

$$E(LOS) = \sum_{t=1}^{K} t * [S_i(t-1) - S_i(t)], \tag{2.4}$$

where K is the expected maximum LOS in days. However, it should be noted that the effect of occupancy on the survival function will vary based on all patient and occupancy level variables. The actual change in LOS will be dependent on the underlying hazard function for the patient given his characteristics. As explained by Hoetker (2007), the treatment directly affects the hazard rate, but that rate is also determined by the patient's other characteristics as well as the baseline level of the variable of interest. For example, a 1% increase in occupancy from 0% occupancy can have a different change in LOS than a 1% increase at 50% occupancy, and the same 1% increase will have a different effect if a patient is 65 years old or 85 years old. As it is difficult to interpret hazard rate changes (the β's in eq. 2.2) (Hoetker 2007, Kuntz et al. 2013), we present the (1) hazard rate as well as (2) an OLS regression for effect sizes. We will use OLS regression to

interpret the magnitude of the change in LOS associated with increased workload, using the following

$$Ln(\mathrm{LOS_i}) = \beta_0 + \beta_1 * \mathrm{Controls_i} + \beta_2 * \mathrm{Occupancy_{i,0}} + \beta_3 * \mathrm{BusyOcc1_{i,0}} + \beta_4 * \mathrm{BusyOcc2_{i,0}} + \beta_5 * \mathrm{Occupancy_{i,d}} + \beta_6 * \mathrm{BusyOcc1_{i,d}} + \beta_7 * \mathrm{BusyOcc2_{i,d}} + \varepsilon_i \quad (2.5)$$

We use Stata 12 for all of our analysis. The hazard functions are solved using maximum likelihood estimation on our inpatient hospital dataset, resulting in a model of defined coefficients for each hazard function.

2.7 RESULTS

Our analyses show consistent results with both the hazard and OLS models. We use the hazard models to test the significance of our variables because those models because it does not force a distribution for the LOS, and is thus a better model for our data (build-up of our cloglog model is shown in **Figure A.1** in the **Appendix**). However, because the coefficients are difficult to interpret, we use OLS models to provide the effect sizes, as well as to compare effect sizes across categories.

2.7.1 OCCUPANCY ON DAY OF ADMISSION

In Hypothesis 1a, we predict that increases in any—including other patient type—workload on day of admission will result in changes in LOS. As shown in **Figure 2.7.1**, on the day of admission, higher occupancy from different patient types is associated with longer LOS, providing strong support for H1a. To illustrate, the Figure 2.7.1 coefficients (shown in bold) in the row "Medical main effect" are negative and significant for all five surgical DRG columns (range from -0.0085 to -0.011; $p<0.01$). The hazard rate coefficient represents the effect on the probability of failure, which in this case is equivalent to leaving the hospital. Therefore, a coefficient less than zero means a patient is less likely to the leave the hospital on a given day, and therefore has a longer LOS than expected based on his or her DRG and unique clinical conditions. The coefficients reported in the row "main effect" of occupancy indicate the effect on the hazard rate of a 1% increase in occupancy for all occupancy levels between 0% and 100%, while the busyness occupancy rows ("busy" and "very busy") correspond to the marginal change in the probability of discharge when occupancy is over the threshold values. The surgical main effect coefficients (also in bold) are negative and significant for all five medical DRG columns

Figure 2.7.1. Complementary log-log regression of the effect of inpatient occupancy on **Day of Admission** on the hazard at time (t).

	Medical Patients					**Surgical Patients**				
DRG:	*871*	*603*	*194*	*291*	*690*	*460*	*481*	*419*	*742*	*470*
Outcome: Hazard at time t h(t):	Septicemia or severe Sepsis w/o MV 96+ hours w/ MCC	Cellulitis w/o MCC	Simple Pneumonia & Pleurisy w/ CC	Heart Failure & Shock w/ MCC	Kidney & Urinary Tract Infections w/o MCC	Spinal Fusion except Cervical w/o MCC	Hip & Femur Procedures except Major Joint w/ CC	Laparoscopic Cholecystec-tomy w/o CDE w/o CC/MCC	Uterine & Adnexa Proc, for non-malig. w/CC/MCC	Major Joint Replace. or Reattach. of Lower Extrem. w/o MCC
Admission Occupancy										
Medical										
Main Effect	-0.0085**	-0.0104**	-0.0108**	-0.0100**	-0.0096**	**-0.0070****	**-0.0095****	**-0.0090****	**-0.0079****	**-0.0099****
	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.003)	(0.001)
Busy (≥85% & <93%)	0.0083**	0.0083**	0.0123**	0.0108**	0.0105**	0.0075	0.0076*	0.0115**	-0.0120	0.0099**
	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)	(0.009)	(0.003)	(0.004)	(0.009)	(0.004)
Very Busy (≥93%)	-0.0019	0.0018	-0.0088*	-0.0083**	-0.0083+	0.0126	0.0016	-0.0121	0.0330*	-0.0013
	(0.002)	(0.003)	(0.004)	(0.003)	(0.004)	(0.018)	(0.007)	(0.010)	(0.016)	(0.008)
Surgical										
Main Effect	**-0.0030****	**-0.0038****	**-0.0031****	**-0.0037****	**-0.0033****	-0.0051**	-0.0038**	-0.0060**	-0.0058**	-0.0053**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.002)	(0.001)	(0.001)	(0.002)	(0.001)
Busy (≥85% & <93%)	0.0001	0.0012	0.0000	0.0035+	0.0015	0.0060	0.0002	0.0011	0.0159+	0.00045
	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)	(0.008)	(0.003)	(0.005)	(0.009)	(0.004)
Very Busy (≥93%)	0.0072*	-0.0022	-0.0025	0.0010	0.0000	-0.0003	-0.0014	0.0115	-0.0498**	0.0044
	(0.003)	(0.004)	(0.004)	(0.004)	(0.005)	(0.022)	(0.007)	(0.011)	(0.019)	(0.008)
Obstetrics										
Main Effect	**-0.0002**	**-0.0010****	**-0.0003**	**-0.0007***	**-0.0013****	**0.0001**	**-0.0002**	**-0.0006**	**-0.0019**	**0.0008**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)	(0.001)	(0.001)	(0.001)
Busy (≥70% & <85%)	0.0000	0.0016+	-0.0005	0.0007	0.0024*	0.0008	-0.0018	0.0004	0.0003	-0.0021
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.005)	(0.002)	(0.002)	(0.004)	(0.002)
Very Busy (≥85%)	0.0000	-0.0036+	0.0021	-0.0009	-0.0031	-0.0042	0.0007	-0.0027	-0.0019	-0.0006
	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)	(0.010)	(0.003)	(0.006)	(0.011)	(0.004)
Discharge Occupancy			Included					Included		
Patient Controls[∓]			Included					Included		
Hospital			Included					Included		
Day of Stay			Included					Included		
Constant	-1.016**	-2.555**	-3.716**	-3.511**	-2.370**	-14.94**	-5.496**	-4.478**	-5.255**	-4.611**
Obs (Patient-days)	403,125	142,946	153,172	190,158	118,890	9,140	49,467	24,722	8,211	46,102

Effects of occupancy on the day of admission on the probability of discharge, by DRG type. A negative coefficient indicates a lower probability of discharge (longer LOS). Other type effects (H1a) are in bold; Own type effects (H1b) are shaded; Occupancies between 0 and 100%; Busy and Very Busy coefficients are marginal effects in the noted occupancy ranges.

[∓]Patient Controls include: age, age^2, day of week, year, race, gender, payment type, and disposition; $^+p<0.1$; $^*p<0.05$; $^{**}p<0.01$; *Standard Errors in parentheses*

(range from -0.0038 to -0.0060; $p<0.01$). Higher obstetrics occupancy, which represents only the effect of shared hospital-wide resources on medical and surgical LOS, provides partial support for H1a with statistical increases in LOS due to obstetric workload for three of the 10 DRGs in the survival model (shown in bold in Figure 2.7.1). It is not unexpected that high surgical (medical) occupancy has a larger effect on the workload of medical (surgical) physicians than obstetrics occupancy does due to more shared resources (e.g., physicians) between surgical and medical patients than between obstetrics and medical (surgical) patients.

To interpret the practical significance of these spillover effects, we use the OLS results shown in **Figure 2.7.2**. Please note that in Figure 2.7.2 we only include OLS coefficients that correspond to significant coefficients in the cloglog model and have the appropriate sign. To calculate the average change in LOS in hours due to a 5% increase in occupancy, for each DRG we first take the exponential of the OLS coefficient multiplied by 5, and then subtract 1 to get the % change in LOS. We then multiply this number by the LOS in days for patients in our dataset with that DRG (see **Figure 2.5.1** for the average LOS for each DRG) and then multiply by 24 hours/ day. For example, for the impact of surgical patient "main effect" occupancy on day of admission on the medical DRG 871, we take the exponential of the main effect coefficient of 0.001 from Figure 2.7.2 multiplied by 5, and then subtract 1. We then multiply this value by the LOS for DRG 871 of 8.85 days * 24 hours/ day = 1.06 hours. Finally, we average the changes in LOS in hours across the five DRGs of that type to get the average. The average increase in LOS of medical patients due to a 5% in surgical patient occupancy on day of admission is 0.6 hours. Similarly, every 1% increase in medical occupancy increases surgical patients' LOS by 0.06% to 0.25%. Translating this effect into hours using the average LOSs for each of the five surgical DRGs, which range from 3.9 to 7.1 days, a 5% increase in medical occupancy is associated with an average increase in surgical LOS of 0.9 hours.

As Figure 2.7.1 shows, Hypothesis 1b is partially supported. The coefficients for own type main effects on the day of admission are all negative and statistically significant. Figure 2.7.2 allows us to compare the effect sizes for own and other type occupancies on LOS. The coefficients for the main effect of day of admission medical occupancy on LOS of medical patients are *more positive*, indicating a stronger impact on LOS, than are the coefficients for surgical (Wald test $p<0.05$ for three of the five DRGs) and obstetrics (Wald test $p<0.05$ for all five DRGs) occupancy. To illustrate, as the shaded row in Figure 2.7.2 shows, for the five medical DRGs, the medical main effect coefficient is

Figure 2.7.2. OLS regression of the effect of inpatient occupancy on **Day of Admission** on LOS

	Medical Patients					**Surgical Patients**				
DRG:	*871*	*603*	*194*	*291*	*690*	*460*	*481*	*419*	*742*	*470*
Outcome: Ln length of stay for:	Septicemia or severe Sepsis w/o MV 96+ hours w/ MCC	Cellulitis w/o MCC	Simple Pneumonia & Pleurisy w/ CC	Heart Failure & Shock w/ MCC	Kidney & Urinary Tract Infections w/o MCC	Spinal Fusion except Cervical w/o MCC	Hip & Femur Proc. except Maj. Joint w/ CC	Laparo-scopic Choley. w/o CDE w/o CC/MCC	Uterine & Adnexa Proc. for non-malig. w/ CC/MCC	Maj. Joint Replace./ Reattach. of Lower Extrem. w/o MCC
Admission Occupancy										
Medical										
Main Effect	0.0029**	0.0013**	0.0017**	0.0032**	0.0010**	**0.0006**	**0.0025****	**-0.0004**	**0.0013**	**0.0008+**
Busy (≥85% & <93%)	-0.0028*	0.0010	-0.0021+	-0.0022	-0.0008	-0.0047	-0.0021	-0.0013	0.0013	-0.0002
Very Busy (≥93%)	0.0005	-0.0041+	0.0048	0.0036	0.0007	0.0018	0.0005	0.0033	-0.0050	-0.0013
Surgical										
Main Effect	**0.0010****	**0.0007****	**0.0004**	**0.0012****	**0.0005***	0.0024*	0.0014**	0.0003	0.0026**	0.0011**
Busy (≥85% & <93%)	0.0019	-0.0009	-0.0006	-0.0012	-0.0006	-0.0109*	0.0007	0.0007	-0.0049	0.0023
Very Busy (≥93%)	-0.0074*	0.0021	0.0055+	-0.0004	-0.0004	0.0257	0.0036	-0.0028	0.0127	-0.0071+
Obstetrics										
Main Effect	**-0.0001**	**0.0005****	**-0.0001**	**0.0005***	**0.0004****	**-0.0007**	**-0.0001**	**0.0001**	**0.0001**	**-0.0004***
Busy (≥70% & <85%)	0.0006	-0.0009+	0.0006	0.0001	-0.0010+	0.0016	0.0007	-0.0006	0.0025	0.0020**
Very Busy (≥85%)	-0.0015	0.0015	-0.0014	-0.0014	0.0012	-0.0017	-0.0004	0.0011	-0.0053	-0.0026
Discharge Occupancy			Included					Included		
Patient Controls[Ŧ]			Included					Included		
Hospital			Included					Included		
Constant	2.944**	1.112**	1.873**	2.194**	1.392**	2.693**	2.409**	1.244**	1.795**	1.531**
Obs (patients)	59,998	44,697	41,711	38,947	42,512	1,930	13,876	13,258	2,579	15,603
R-Squared	0.314	0.157	0.14	0.255	0.158	0.392	0.209	0.156	0.364	0.226

This table shows the effects of occupancy (by type) on the day of admission on the LOS for patients by DRG type, where a positive coefficient indicates a longer LOS. These coefficients are used to calculate the magnitude of the occupancy effects. Other type effects (H1a) are in bold; Own type effects (H1b) are shaded; Occupancies are between 0 and 100%; Busy and Very Busy coefficients represent marginal effects in the noted occupancy ranges. [Ŧ]Patient Controls include: age, age^2, day of week, year, race, gender, payment type, and disposition; $^+p<0.1$; $^*p<0.05$; $^{**}p<0.01$

significant and has a greater magnitude than the surgical main effect coefficient and the obstetric coefficient (both in bold). Similarly, for the five surgical DRGs, the surgical main effect coefficient is significant and of greater magnitude (shaded) than the obstetric coefficient (bold) (Wald test $p<0.05$ for four of the five DRGs). These results provide support for H1b. However, as Figure 2.7.2 shows, H1b is only partially supported because the coefficients for surgical occupancy on day of admission on surgical LOS are not all larger than the medical main effect coefficients for the surgical DRGs. The differences are statistically significant for only two of the five DRGs (Wald test $p<0.01$).

As shown in Figure 2.7.2, a 1% increase in medical occupancy up to 85% is associated with between a 0.1% and 0.32% increase in medical patients' LOS. In our sample, this equates to an average increase of 1.7 additional hours in the hospital for every 5% increase in medical patient occupancy up to 85% on the day of a patient's arrival. The effect of surgical occupancy on medical patient LOS is significantly lower than the effects of medical occupancy on medical LOS, with own type occupancy having a 2.7 times greater effect on LOS. The difference between the effect of increased medical and increased surgical occupancy on medical patient LOS (approximately 1.1 hours for a 5% increase in occupancy), represents the effects of workload smoothing because the impact of surgical patients on medical patient LOS is primarily due to congestion effects. In contrast, the impact of medical patient occupancy on medical patient LOS is from both congestion effects and workload smoothing as medical clinicians can smooth their own workload by discharging medical patients earlier. Using similar calculations and coefficients from Figure 2.7.2, a 1% increase in surgical occupancy at all occupancy levels is associated with an increase in surgical LOS from 0.03% to 0.26%. Equivalently, a 5% increase in surgical occupancy results in an average increase in LOS of 1.0 hours. Thus, medical patient occupancy's impact on surgical LOS is approximately 90% of the corresponding effect of surgical occupancy on surgical LOS. These results provide support that higher occupancy on the day of admission is associated with longer LOSs, with own type occupancy having a greater effect than other type occupancy.

2.7.2 OCCUPANCY ON DAY OF DISCHARGE

Figure 2.7.3 shows the results from our test of H2a and H2b. H2a predicted a U-shaped effect on LOS of other type of occupancy at the start of the day of discharge. As shown in Figure 2.7.3, the coefficients in bold show that this hypothesis is supported. The coefficient for the main effect of surgical occupancy on day of discharge is positive and significant for all five medical DRGs ($p<0.01$). This indicates that as occupancy increases up to 85%, LOS gets shorter, which is consistent with the downward slope of the U-shape. The coefficient for main effect of medical occupancy on surgical LOS is also positive and significant for all five surgical DRGs ($p<0.01$). Furthermore, the main effect of obstetric occupancy is positive and significant for all 10 DRGs ($p<0.01$). Thus, there is strong support for the downward slope of the U-shape. The upward slope is also supported. All ten DRGs have a negative, significant coefficient for the "busy" effect of

Figure 2.7.3. Complementary log-log regression of the effect of inpatient occupancy on **Day of Discharge** on the hazard at time (t).

	Medical Patients					Surgical Patients				
DRG:	*871*	*603*	*194*	*291*	*690*	*460*	*481*	*419*	*742*	*470*
Outcome: Hazard at time t h(t):	Septicemia or severe Sepsis w/o MV 96+ hours w/ MCC	Cellulitis w/o MCC	Simple Pneumonia & Pleurisy w/ CC	Heart Failure & Shock w/ MCC	Kidney & Urinary Tract Infections w/o MCC	Spinal Fusion except Cervical w/o MCC	Hip & Femur Procedures except Major Joint w/ CC	Laparoscopic Cholecystec-tomy w/o CDE w/o CC/MCC	Uterine & Adnexa Proc, for non-malig. w/ CC/MCC	Major Joint Replace. or Reattach. of Lower Extrem. w/o MCC
Discharge Occupancy										
Medical										
Main Effect	0.0381**	0.0340**	0.0351**	0.0374**	0.0328**	**0.0284****	**0.0307****	**0.0250****	**0.0292****	**0.0245****
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.002)	(0.001)
Busy (≥85% & <93%)	-0.0845**	-0.0738**	-0.0761**	-0.0806**	-0.0693**	**-0.0623****	**-0.0672****	**-0.0518****	**-0.0603****	**-0.0539****
	(0.001)	(0.002)	(0.001)	(0.001)	(0.002)	(0.011)	(0.003)	(0.004)	(0.009)	(0.003)
Very Busy (≥93%)	0.0657**	0.0586**	0.0627**	0.0634**	0.0531**	**0.0276**	**0.0539****	**0.0446****	**0.0443***	**0.0450****
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.021)	(0.007)	(0.009)	(0.019)	(0.008)
Surgical										
Main Effect	**0.0134****	**0.0123****	**0.0122****	**0.0129****	**0.0113****	0.0213**	0.0150**	0.0132**	0.0175**	0.0160**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.002)	(0.001)	(0.001)	(0.002)	(0.001)
Busy (≥85% & <93%)	**-0.0399****	**-0.0323****	**-0.0328****	**-0.0383****	**-0.0346****	-0.0490**	-0.0358**	-0.0340**	-0.0356**	-0.0351**
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.009)	(0.003)	(0.005)	(0.008)	(0.004)
Very Busy (≥93%)	**0.0335****	**0.0272****	**0.0285****	**0.0356****	**0.0312****	0.0406*	0.0284**	0.0267*	0.0069	0.0365**
	(0.003)	(0.004)	(0.004)	(0.004)	(0.005)	(0.017)	(0.007)	(0.010)	(0.019)	(0.010)
Obstetrics										
Main Effect	**0.0081****	**0.0071****	**0.0068****	**0.0075****	**0.0064****	**0.0081****	**0.0057****	**0.0065****	**0.0084****	**0.0072****
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)
Busy (≥70% & <85%)	**-0.0179****	**-0.0149****	**-0.0143****	**-0.0167****	**-0.0119****	**-0.0133****	**-0.0127****	**-0.0148****	**-0.0177****	**-0.0114****
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.005)	(0.002)	(0.002)	(0.004)	(0.002)
Very Busy (≥85%)	**0.0118****	**0.0096****	**0.0086****	**0.0100****	**0.0054****	**0.0016**	**0.0063**	**0.0131****	**0.0231***	**0.0036**
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.010)	(0.004)	(0.004)	(0.009)	(0.004)
Admission Occupancy			Included					Included		
Patient Controls[∓]			Included					Included		
Hospital			Included					Included		
Day of Stay			Included					Included		
Constant	-1.016**	-2.555**	-3.716**	-3.511**	-2.370**	-14.94**	-5.496**	-4.478**	-5.255**	-4.611**
Obs (Patient-days)	403,125	142,946	153,172	190,158	118,890	9,140	49,467	24,722	8,211	46,102

Effects of occupancy on the day of discharge on the probability of discharge, by DRG type. A negative coefficient indicates a lower probability of discharge (longer LOS). Other type effects (H1a) are in bold; Own type effects (H1b) are shaded; Occupancies between 0 and 100%; Busy and Very Busy coefficients are marginal effects in the noted occupancy ranges.

[∓]Patient Controls include: age, age^2, day of week, year, race, gender, payment type, and disposition; $^+p<0.1$; $^*p<0.05$; $^{**}p<0.01$; *Standard Errors in parentheses*

surgical, medical, and obstetric occupancy on the day of discharge ($p<0.01$). It should be noted that the spline effects are marginal, so all coefficients (main and busy) must be added together to get the net effect of "busy" occupancy. Therefore, as occupancy of medical (surgical) or obstetrics patients on the day of discharges increases in the range between 85% and 93% of maximum occupancy, surgical (medical) patients' LOSs increases. Finally, we test the impact of "very busy" (>=93%) occupancy. As with the "busy" effect above, the net effect of an increase in occupancy in the "very busy" range is the summation of all coefficients (main, busy, and very busy). For very busy surgical occupancy on day of discharge, LOS flattens or decreases for all five medical DRGS ($p<0.01$). Similarly, very busy medical occupancy flattens or decreases LOS for four of the five surgical DRGs ($p<0.05$). Very busy (>=85%) obstetric occupancy flattens or decreases LOS for all five medical DRGS ($p<0.01$) and two surgical DRGs ($p<0.05$).

Same type occupancy also has a U-shaped impact on LOS, and the slopes are steeper than other type occupancy for each portion for medical patients, providing partial support for H2b. As Figure 2.7.3 shows, increases in medical (surgical), occupancy, up to the busy range, is associated with shorter LOSs for all medical (surgical) patients ($p<0.01$), consistent with the U-shaped effect. In addition, at busy occupancy levels, the LOS increases as occupancy increases ($p<0.01$). At very busy occupancy levels (>=93%), the LOS flattens or decreases slightly ($p<0.05$) for all but one DRG. Again, because the spline effects are marginal, the main effect and busy coefficients must be added together to get the net busy effect, while all of the coefficients (main effect, busy effects, and very busy effects) must be added together to get the net effect in the very busy range. Therefore, while all of the coefficients are negative at the very busy occupancy levels, the total magnitude of the sum of the effects is approximately 0 for some of the DRGs.

Figure 2.7.4 shows the results from OLS regression. We use those coefficients to quantify the impact of day of discharge occupancy on LOS using the same method as described in Section 2.7.1, as shown in the notes of **Figure 2.7.5**, which provides a summary of the impact of a 5% change in occupancy at the various occupancy levels on LOS in hours. We first describe the impact of other type occupancy on LOS. A 5% increase in surgical occupancy on day of discharge has the following average impact on LOS of medical patients: (1) it decreases LOS by 1.5 hours up until occupancy is 85%; (2) it increases LOS by 3.8 hours between 85% to 93% occupancy; (3) and over 93% occupancy, it decreases LOS by 1.6 hours. Similarly, as medical occupancy on day of discharge increases by 5%, it has the following average impact on LOS of surgical

patients: (1) up to 85% occupancy, it decreases LOS by 2.0 hours; (2) between 85% to 93%, it increases LOS by 2.5 hours; (3) over 93%, it decreases LOS for the statistically significant DRGs by an average of 2.6 hours. Finally, as obstetrics occupancy increases 5% on day of discharge, while remaining below the busy threshold of 70%, medical LOS decreases medical by an average of 1.1 hours and surgical LOS by an average of 0.8 hours.

Figure 2.74. OLS regression of the effect of inpatient occupancy on **Day of Discharge** on LOS

	Medical Patients					**Surgical Patients**				
DRG:	*871*	*603*	*194*	*291*	*690*	*460*	*481*	*419*	*742*	*470*
Outcome: Ln length of stay for:	Septicemia or severe Sepsis w/o MV 96+ hours w/ MCC	Cellulitis w/o MCC	Simple Pneumonia & Pleurisy w/ CC	Heart Failure & Shock w/ MCC	Kidney & Urinary Tract Infections w/o MCC	Spinal Fusion except Cervical w/o MCC	Hip & Femur Proc. except Maj. Joint w/ CC	Laparo-scopic Choley. w/o CDE w/o CC/MCC	Uterine & Adnexa Proc. for non-malig. w/ CC/MCC	Maj. Joint Replace./ Reattach. of Lower Extrem. w/o MCC
Discharge Occupancy										
Medical										
Main Effect	-0.0077**	-0.0023**	-0.0029**	-0.0071**	-0.0021**	**-0.0010**	**-0.0053****	**0.0003**	**-0.0047****	**-0.0011+**
Busy (≥85% & <93%)	0.0182**	0.0073**	0.0086**	0.0157**	0.0054**	**-0.0009**	**0.0120****	**0.0026+**	**0.0093+**	**0.0014**
Very Busy (≥93%)	-0.0144**	-0.0073**	-0.0117**	-0.0154**	-0.0033	**0.0224+**	**-0.0118****	**-0.0041**	**0.0003**	**0.0037**
Surgical										
Main Effect	**-0.0030****	**-0.0010****	**-0.0011****	**-0.0029****	**-0.0010****	-0.0039**	-0.0036**	-0.0002	-0.0043**	-0.0022**
Busy (≥85% & <93%)	**0.0114****	**0.0028***	**0.0022**	**0.0092****	**0.0058****	0.0074	0.0059**	0.0007	0.0079	0.0024
Very Busy (≥93%)	**-0.0114****	**-0.0026**	**-0.0018**	**-0.0097***	**-0.0065***	-0.0082	-0.0043	0.0018	-0.0002	-0.0020
Obstetrics										
Main Effect	**-0.0019****	**-0.0012****	**-0.0009****	**-0.0022****	**-0.0008****	**-0.0026****	**-0.0012****	**-0.0002**	**-0.0011**	**-0.0011****
Busy (≥70% & <85%)	**0.0030****	**0.0017****	**0.0008**	**0.0046****	**0.0003**	**0.0061***	**0.00209+**	**0.0005**	**0.0026**	**0.0001**
Very Busy (≥85%)	**0.0001**	**-0.0018**	**0.0001**	**-0.0025+**	**0.0012**	**-0.0051**	**0.0023**	**-0.0006**	**-0.0062**	**0.0021**
Admission Occupancy			Included					Included		
Patient Controls[Ŧ]			Included					Included		
Hospital			Included					Included		
Constant	2.944**	1.112**	1.873**	2.194**	1.392**	2.693**	2.409**	1.244**	1.795**	1.531**
Obs (patients)	59,998	44,697	41,711	38,947	42,512	1,930	13,876	13,258	2,579	15,603
R-Squared	0.314	0.157	0.14	0.255	0.158	0.392	0.209	0.156	0.364	0.226

This table shows the effects of occupancy (by type) on the day of discharge on the LOS for patients by DRG type, where a positive coefficient indicates a longer LOS. These coefficients are used to calculate the magnitude of the occupancy effects. Other type effects (H1a) are in bold; Own type effects (H1b) are shaded; Occupancies are between 0 and 100%; Busy and Very Busy coefficients represent marginal effects in the noted occupancy ranges.
[Ŧ]Patient Controls include: age, age^2, day of week, year, race, gender, payment type, and disposition
$^+$p<0.1; *p<0.05; **p<0.01

In the busy range (i.e. ›70% occupancy and ‹85%), the direction of the effect reverses so that a 5% increase in occupancy increases the LOS of both medical and surgical patients

by an average of 0.6 hours. In the very busy range (i.e. ›85% occupancy), a 5% increase in occupancy results in a slight decrease in LOS for all medical patients of 0.5 hours, and a 1.2 hour decrease in LOS for the two of the five surgical DRGs that were statistically significant. These results are consistent with clinicians responding to increasing occupancy by smoothing their workload by discharging patients who are close to the end of their stay. However, at high workload (between 85 and 93% occupancy), congestion effects dominate, resulting in longer LOS. At very high levels of occupancy, LOS again decreases, which we discuss in Section 2.8 below.

Figure 2.7.5. Average impact of a 5% change in **Day of Discharge** occupancy on average LOS in hours

Occupancy Type	**Type**	**Main Effect[1]**	**Busy[2]**	**Very Busy[3]**
Medical on Medical	Same	-3.7	5.6	-3.0
Surgical on Medical	Other	-1.5	3.8	-1.6
Obstetrics on Medical	Other	-1.1	0.6	-0.5
Surgical on Surgical	Same	-1.8	1.3	-1.9
Medical on Surgical	Other	-2.0	2.5	-2.6
Obstetrics on Surgical	Other	-0.8	0.6	-1.2

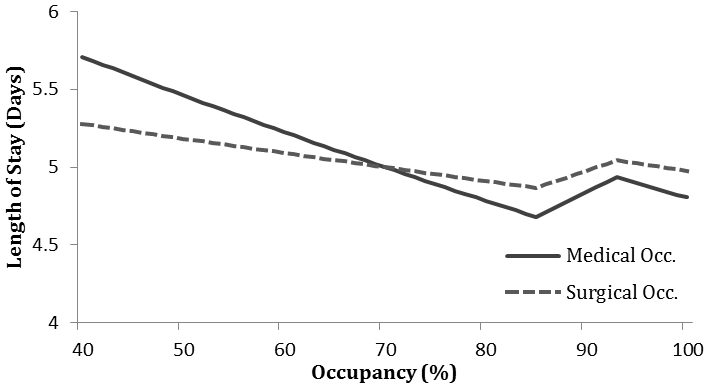
[1](exp(coefficient*10)–1)*(DRG avg. LOS in days)*24 hours/day. Averaged across all significant DRGs in Figure 2.7.3 with OLS coefficients with the right sign in Figure 2.7.4

[2](exp((coefficient main+coefficient busy)*10)-1)*(DRG avg. LOS in days)*24. Averaged across all significant DRGs.

[3](exp((coefficient main+coefficient busy+coefficient very busy)*10)-1)*(DRG avg. LOS in days)*24. Averaged across all significant DRGs.

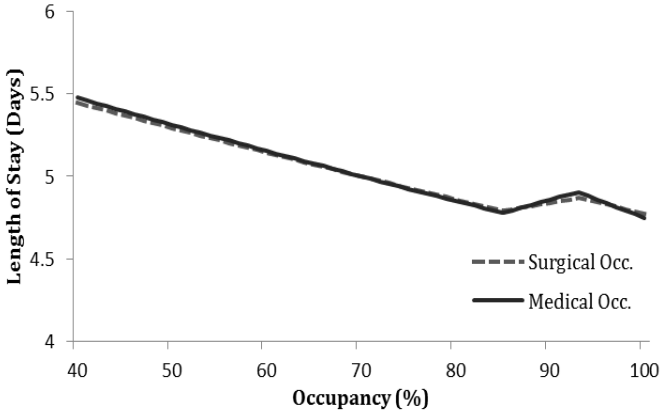
Next, we describe the impact of same type occupancy. As medical occupancy on day of discharge increases by 5% (up to 85% occupancy), the LOS of medical patients decreases on average by 3.7 hours. In the busy range, between 85% and 93%, the same increase in occupancy results in an average increase in LOS of 5.6 hours, while in the very busy range, i.e. › 93% occupancy, the effect is a 3.0 hour reduction in LOS. For a 5% increase in surgical occupancy up to 85% occupancy, the average reduction in LOS for surgical DRGs is 1.8 hours. As surgical occupancy enters the busy range of 85% to 93%, a 5% increase in occupancy results in an average net increase in the LOS of surgical patients of 1.3 hours. In the very busy range, above 93%, there is an average reduction in LOS of 1.9 hours. As shown in **Figure 2.7.5,** the effects of obstetrics occupancy are smaller, on average, than for own type occupancy for all patients, and surgical occupancy has a smaller effect than medical occupancy on medical patient LOS, providing partial support for H2b. **Figures 2.7.6 and 2.7.7** graphically show these results for medical and surgical patients, respectively.

Figure 2.7.6. Discharge Occupancy on Medical Patient LOS.



The *solid* (dashed) line shows the average change in LOS due to a change in *medical* (surgical) occupancy, holding all else equal, of a medical patient with an expected LOS of 5 days at 70% *medical* (surgical) occupancy.

Figure 2.7.7. Discharge Occupancy on Surgical Patient LOS.



The *solid* (dashed) line shows the average change in LOS due to a change in *medical occupancy* (surgical occupancy), holding everything else equal, of a surgical patient with an expected LOS of 5 days given a 70% *medical occupancy* (surgical occupancy) level.

2.7.3 CLINICIAN EXPERIENCES

Discussions with physicians and nurses about admission and discharge processes provide additional support for our hypotheses. First, with regard to the increased LOS due to high occupancy of other type inpatients on the day of admission, one physician told us that patients will have a longer length of stay because "everybody will be fighting for beds, competing for space, so our patients may get stuck in the ED." Moreover, the physician said that "if you don't get what you need done on [your patient's] first day, or something is missing, then the patient is usually not going to progress fast enough and

you basically lost a day." Our interviews with nurses provide further explanation as to why a newly admitted patient will have a longer length of stay if admitted on a day when the inpatient unit is busy. All five nurses whom we interviewed stated that admissions are time consuming, and as a result, they try to first make sure that their current patients' needs are met before beginning the process of admitting a new patient. This results in a delay to the start of care for the newly admitted patient. For example, one nurse described the challenge a co-worker had trying to balance the work demands of a new admission with her existing patients. "She was struggling to juggle her patients, 'Should I go see this new patient first? But if I see this new patient, I'm missing the opportunity to see my other three patients.' She decided to see the new admission after caring for her existing patients, a decision that was somewhat justified because the nurse reported to us that "when she did attend to the newly admitted patient, she was in that room for at least a good hour." As a result of this dynamic, we find that nurses and physicians who get a new admit will often first care for all their existing patients before working on the newly admitted patient.

Around the time of discharge, the clinicians mentioned competing effects as workload increased. In particular there is "…the incentive to work hard discharging patients is [to] have fewer people the next day." In one hospital where we interviewed, hospital leadership also puts pressure to discharge patients. A physician told us that "when the hospital reaches a certain capacity, the administrators send out texts or constantly berating everyone to hurry up and discharge so that new patients can come in." However, a physician noted that when it gets very busy, "discharges, getting orders in…get delayed because I'm not physically able to input [them]." Thus, the clinician interviews confirm the u-shaped relationship our data show between increasing workload on day of discharge and LOS.

We also asked physicians about spillover effects, such as their perception of their patients' treatment times when the hospital, though not necessarily their own inpatient unit, gets busier. A physician with administrative responsibilities who is thus aware of the hospital as a whole, explained "if there are a lot of patients in the hospital, everyone is fighting for resources – there's imaging or lab tests or even just people, manpower." These responses are consistent with our empirical results, and we will discuss the implications in section 2.8.

2.7.4 Robustness

One of the assumptions for cloglog analysis is proportional hazards. Specifically, the effect on the hazard rate does not change over the course of a patient's stay. To test this assumption, we run an analysis interacting all variables with the day of the stay, and then perform Wald tests for each of our variables of interest. There are no consistently statistically significant variables across all of the DRG models, so we feel comfortable with the proportional hazards assumption.

As mentioned in Section 2.5, we also ran robustness tests on our "busy" and occupancy definitions. In our analyses we define busy occupancy levels in order to perform a spline analysis on our data. While this technique allows us to see changes in the effects of occupancy at different levels, it also requires us to make a choice as to what we consider "busy" and "very busy" which we base off of the literature and discussions with practitioners, since there is no standard definition. As a robustness check, we perform a sensitivity analysis (using our OLS models for ease of comparison) on our definition of busy, testing the effects of up to a +/-5% change (in increments of 1%) in the definition of busy. We do the same for the definition of very busy. We find that the direction and order of magnitude of our coefficients of interest are not very sensitive to our definitions of busy and very busy.

To test our occupancy calculations, we calculate an occupancy measure that incorporates a possible weekend effect since some hospitals schedule surgeries early during a week to have fewer patients on weekends (KC and Terwiesch 2014), and thus may have more limited services on the weekend so that the maximum number of patients that can be seen is lower than on weekdays. We account for this by modifying our occupancy calculation. Previously, we calculated a rolling maximum (99^{th} percentile) over the previous 90 days. As a robustness check, we categorize each day into a weekday or weekend and taking the rolling maximum (99^{th} percentile) by type for the previous 90 days. For example, to calculate the occupancy for hospital "H" on Sunday, June 8, 2008, we first find the 99^{th} percentile of the number of patients who were inpatient on a weekend in hospital H from March 10, 2008 through June 8, 2008, and label that the maximum. We then divide the number of patients actually in the hospital on June 8 by that maximum. We analyze the data using the same cloglog function. In addition, we test that our results are not sensitive to using a 90 day rolling maximum as a proxy for the hospital beds available. As the demand for inpatient beds is often seasonal, we perform an additional test using the maximum patient census in each quarter of the year as the

denominator for our occupancy calculation. In both robustness tests the results are consistent with those found with our original definition of occupancy.

Related to the weekend effect in which there could be a difference in available resources, we also run an analysis where we restrict our sample to patients admitted on weekdays. In addition, we test the robustness of our definition of LOS. We define LOS for surgical patients as the discharge date minus the day of surgery for surgical patients. We do the latter since when a surgery is performed may be a function of the time or day of the week that a patient arrives at the hospital, such that a patient who arrives in the evening may not be able to have surgery until the following morning. Therefore, the true care begins after the patient has had his surgery. However, to ensure that using this measure of LOS does not affect our results, we run a robustness test for surgical patients using the day of arrival to calculate the LOS used in the cloglog model. Next, we test whether our results are robust to our standard error clustering. Instead of clustering by hospital, we cluster by patient (each patient has multiple observations since an observation is a patient-day in the cloglog model). In summary, all of our robustness check results are consistent with our main model.

Finally, we measure the effects of workload at the start and end of a patient's stay because we assumed based on discussions with clinicians that these are the times when the majority of decisions are made. We test this assumption by analyzing the effects of occupancy on other days. However, given the high correlation in occupancy across days, we limit our sample to DRGs with longer average LOSs (DRGs 603, 194, 481, and 471). Within these DRGs, we use only those patients who stayed in the hospital for at least 7 days. This enables us to leave at least two days between occupancy measures. In addition to the day of admission and day of discharge, we include the occupancy on day 4 (about half way through the stay) in our OLS model. While the effects of workload on the day of discharge remain the same, day of admission effects become mostly insignificant. However, the day four effects are similar to the day of admission effects in our main model. Thus, these results suggest that high occupancy during the first half of a patient's stay is associated with increased LOS due to congestion. For ease of analysis and generalizability for short LOS DRGs, using the first day as a measure is appropriate.

2.8 DISCUSSION AND CONCLUSION

Our work shows that the type and timing, as well as the level, of patient workload matters in terms of LOS, with congestion effects (which increase LOS) dominating

during the initial stage of a patient's stay, in addition to periods of very high load (>85% occupancy) towards the end of a patient's stay. Workload smoothing (which decreases LOS) characterizes the effects of increased load towards the end of the stay at lower than 85% occupancy. Additionally, we show that when resources are shared across separate streams of patients, the congestion and workload smoothing effects "spillover" across patient types, however the spillover effects are smaller than the effects of own type occupancy. In this way, we are able to disentangle the effects that result from the congestion associated with high demand and the effects of worker behavior to smooth their workloads. In doing so, we are able to quantify their relative magnitudes. These results are significant because they show that workload has different effects throughout a patient's stay, which has implications for improvement.

Our results provide support for our hypotheses that both congestion and workload smoothing effects occur in hospitals and spillover across patient types. Moreover, the effects of workload smoothing affect not just the patients being discharged, but also newly admitted patients. Specifically, our findings indicate that when hospitals have high occupancy levels, even when they are not 100% full, clinicians prioritize discharging patients nearing the completion of their stay, which results in newly admitted patients having longer LOSs than they would otherwise. In fact, at the hospital visited by the first author, she observed the chief medical officer walk through each unit of the hospital telling all clinicians to try to move patients through more quickly due to the high occupancy within parts of the hospital. In our analysis, this behavior is evident in that obstetrics occupancy is more likely to affect medical and surgical patients' LOS at the end of their stay than at the beginning. As a result in the delay in treatment at the early part of their stay, the LOS of newly admitted patients increases. We find that at the start of a medical patient's stay, for every 5% increase in own or other type occupancy, there is on average a 0.4% increase in patient LOS (or 0.6 hours) due to congestion, and an additional 0.6% increase in LOS (or 1.1 hours) due to the prioritization of patient discharges as own type occupancy increases 5%. At the end of a patient's stay, an increase in own type occupancy of 5%, up to 85% occupancy, results in an average decrease in LOS for all patients of 1.8% (or 2.7 hours). Above 85% (up to 93%, where the effect levels off or slightly decreases again likely due to the inpatient units being full), each 5% increase in occupancy increases the LOS 2.2 % (or 3.4 hours). As with the congestion effect, this workload smoothing effect spills over across patient types, though the effects is smaller than the same type effect.

At first glance, our finding that LOS increases on the day of admission and at occupancies above 85% on the day of discharge seems to contradict prior literature, which has found shorter LOSs as occupancy increases. However, these results are not necessarily inconsistent because prior studies have focused on settings where the ICU is at or very near 100% occupancy near the time of a patient's discharge, so those employees have no option but to discharge a patient from the ICU to make room for a severely ill incoming patient. In fact, we find at very high levels of occupancy (>93% for medical and surgical occupancy), there appears to be a last push to discharge patients to free up beds.

These results have significant implications for hospital managers and for policy makers. First, we show that there are congestion effects in hospitals related to queuing for shared resources, particularly at the start of a patient's stay. Previous work has only analytically described the queuing effects within hospitals, and how it relates to beds (Green and Nguyen 2001), or focused on early discharge at the end of a patient's stay due to very high patient load (e.g., KC and Terwiesch 2012). Our results suggest that available inpatient beds may not always be the bottleneck resource, as is generally assumed, since congestion effects occur even when the hospital is significantly below maximum capacity. Instead, our results show that shared auxiliary resources, such as radiation, transport, and pharmacy services, also play an important role in LOS. Applying managerial improvements suggested by queuing theory, it is necessary to explore all shared resources within hospitals to determine which are the most likely to be blocked due to high demand. For example, in the California hospital the first author visited, one of the biggest delays for new patients was waiting for transport services to move admitted patients from the ED to the inpatient unit. In other hospitals, the bottleneck may be radiology or pharmacy services. In addition, the capacity for each of these services must be enough to meet the needs of all inpatient types since high occupancy of one patient type affects other patient types. Currently, California has mandates governing the number of beds and nurses per bed that must be met by hospitals, thus it is not a stretch to imagine having similar laws for pharmacists, technicians, and other care providers.

Disentangling queuing from workload smoothing effects also has major managerial implications. Early discharge of patients makes sense under two conditions. The first one is if the hospital is full and a new patient must be admitted. Under these conditions, the incoming patient is sicker than the patients already in the hospital thus has priority over the patient closest to being able to be discharged. Consistent with this scenario, our

results show that at very high occupancies there seems to be an emphasis on discharging patients early. The second condition is if the amount of time saved with an early discharge is greater than increased probability of readmission and the resulting LOS of that readmission. In this situation, the hospital does not need to be full, but could be getting busier, and the benefits of the early discharge could be great enough to justify it. There is recent work (Chan et al. 2012) that analytically models this decision and the point at which early discharge is optimal. However, that paper does not account for the effect that workload smoothing has on other patients in the inpatient unit. By identifying that the prioritization of early discharges over admissions increases the LOS of newly admitted patients, our work suggests that current models underestimate the effects of workload smoothing. In addition, these models assume early discharges only occur when the inpatient unit is full or nearly full, and thus solutions are most concerned with capacity planning. However, we find that clinicians make early discharge decisions even when a unit is not full, which requires a different set of solutions to address these behavioral responses in addition to capacity concerns.

We provide a numerical example to show the impact of underestimating the consequences of workload smoothing technique of prioritizing discharges. We use medical patients as our example because they most clearly account for the impact of discharge prioritization. In California, the average readmission has a LOS 48% longer than a one-time visit (California Office of Statewide Health Planning and Development 2010). Applying these statistics to our sample, the average LOS for a readmission would be 9.4 days, or approximately 225 hours. Based on our results, a 5% increase in occupancy for a patient near discharge decreases the patient's LOS by 3.7 hours. Under current assumptions in the literature, discharging the patient 3.7 hours earlier will save bed-hours if the increased probability of readmission due to the early discharge, multiplied by the expected LOS should the patient be readmitted (which we estimate at 225 hours), is less than 2.7 hours. Thus we have:

$$225 \text{ hours} * p \leq 3.7 \text{ hours} \tag{2.6}$$

where p is the increased probability of being readmitted due to being discharged 3.7 hours early. Solving for p, as long as the early discharge does not increase the probability of being readmitted by more than 1.6%, the early discharge would minimize bed hours used. However, our work shows that while the dischargeable patient has a LOS 3.7 hours shorter, there is a cost for each newly admitted patient in the form of an increase in LOS

of 1.1 hours due to prioritization of the dischargeable patient. Therefore, the net savings in bed-hours is only 2.6 hours. Replacing 3.7 hours with 2.6 hours in equation (2.6) above, we find if the early discharge increases the probability of readmission by 1.1% or more, it is no longer beneficial to discharge the patient early. While the readmission rates vary by clinical condition and hospital, previous work on readmission rate changes associated with higher occupancy suggests that the increase will likely be higher than 1.1% with a shortened LOS (Anderson et al. 2012).

In order to minimize bed hours used, we could imagine a benefit if hospitals can prioritize new patients over discharging currently convalescing patients. While less time will be saved from early discharges, newly admitted patients will have shorter delays in the start of care. Moreover, the reduction in early discharges should reduce the probability of readmission since the discharged patient will have additional convalescing time. Using the same calculations as above, if newly admitted patients who enter when the hospital is busy had an increase in LOS of only 0.6 hours, instead of 1.1 hours, while early discharge times were reduced by that half hour, so that they now were discharged only 3.2 hours early, then the net effect would be the same. However, since the discharged patients have additional time to recover, they are less likely to be readmitted, and may now satisfy the condition that the probability of readmission is less than or equal to 1.1%. We leave it to future research to determine the exact changes in time that are feasible while maximizing capacity utilization.

For hospitals to be able to shift their response to heavy workload from patient discharge to instead the timely start of treatment for newly admitted patients, they will need to develop new capabilities. One possible mechanism to control admission and discharge times to employ admission and discharge nurses and physicians. Both physicians and nurses control the timing of admissions and discharges. While the physicians make the decisions and must write the orders, nurses influence when those orders are completed. The time consuming nature of discharges and admissions suggests that it could be beneficial to have dedicated admission/ discharge nurses and physicians to ensure that this important work is not compromised when staff nurses and physicians become inundated with high workloads from existing patients. It should be noted that we observe the workload smoothing effect even when inpatient units are not at maximum capacity, indicating that these early discharges are not necessary to free up a bed for a new patient. Furthermore, dedicated discharge and admission clinicians may be better able to schedule their tasks to meet demand because they are free from routine

patient care. Finally, if clinicians are assigned to perform specific tasks (such as admissions and discharges), it would be easier to implement more standardized procedures and checklists to remove additional inefficient discretionary behavior. By removing the parts of care which generally require the largest blocks of time, other nurses may be able to attend to more patients, reducing the negative effects of high load. Thus, specialization may prove useful in the goal of reducing the overall LOS of patients.

The exact benefits of these changes will depend on the specific hospital. We analyzed more than 200 hospitals, and some may already be better at managing high workloads. Also, due to data limitations, we cannot specifically test for changes in outcome quality associated with occupancy-related changes in LOS. Thus, we leave it to future work to determine which hospitals perform well under high load, what characteristics define them, and if it results in better clinical outcomes. Finally, we have only studied these effects in the hospital setting. While we think it is highly probable that these results apply to other settings with high worker discretion, such as law firms and call centers, more research is needed to confirm.

There remains much to be explored in this area. We believe that understanding how workload affects LOS and resource use across all hospital inpatient units and patients' entire hospital stays is an important contribution to theory, practice, and policy. Our results show that true optimization of resource use in healthcare organizations comes from exploring not just individual units, but the organization as a whole.

J. Berry Jaeker, Tucker, 2014. Priority and Predictability.
Working Paper. Harvard Business School. Boston, MA.

4

Priority and Predictability

4.1 INTRODUCTION

As the cost of healthcare, and in particular, hospital care, continues to rise in the US and other developed countries, the need for the efficient delivery of care becomes ever more important (Green 2012). The efficient delivery of care is defined as providing the right care to the right patient at the right time. Therefore, hospitals need a system to ensure a patient receives the proper care and there is enough capacity so that resources are available for the care to be dispensed in a timely manner. Generally, all patients could receive timely, appropriate care if the resources in a hospital were limitless. However, there are capacity constraints, so there has been an increased interest in identifying and applying operations management solutions to improve the delivery of hospital care given these constraints.

Improving hospital care delivery is often complicated due to the high degree of variation in the demand over time (Eddy 1984, Green and Nguyen 2001). In addition, traditional operations management theories assume the completion time of tasks is a random variable that is independent of the demand in the system. More recent work shows that in service settings where workers have high levels of discretion, and particularly in healthcare, this assumption does not hold, and patient care is affected by the workload in a hospital (e.g., Berry Jaeker and Tucker 2014, KC and Terwiesch 2012, Kim et al. 2012, Kuntz et al. 2013). Specifically, this stream of research has found that the length of stay (LOS) of a patient, and the probability of admission for an emergency patient, is affected by the occupancy of the hospital. Most of this work takes into account the differences patients' clinical conditions, but there is a more fundamental difference in patients that has been largely unexplored: the admission type of the patient. Patients who are admitted to a hospital can either be emergent (unscheduled) or scheduled, where emergent patients can be emergency medical (EM) patients or emergency surgical (ES), while scheduled patients are nearly always scheduled surgical (SS). There has been some work on the capacity allocation of operating rooms, beds, and staffing for these different admission types, primarily across surgical patients, but to the best of our knowledge, none on the how the variation in admission type affects the LOS of convalescing inpatients during periods of high workload, nor how the workload in the hospital affects the probability of admission for each type.

The goal of this work is to show how the admission type of incoming patients (scheduled or emergent) impacts the care these and other hospital patients receive, particularly under periods of high workload. Using this information, we can identify and quantify managerial levers that can affect patient care as well as offer potential solutions to improve the delivery of care. Specifically, using two years of patient level data from 169 California hospitals, we find that when a surgical inpatient unit within a hospital is nearly full, SS patients are less likely to be admitted, consistent with surgery cancellations as the hospital becomes full. For a hospital with between 30 and 40 beds, and 5 expected SS admission, if the number of available surgical beds, at the start of the day, goes from 5 to 4, the expected number of SS admissions falls to 4.8 patients. However, before a SS is cancelled, the hospital will first attempt to discharge patients early. For the same size hospital with an expected number of surgical discharges of 10, increasing the number of SS patients admitted from 5 to 6, increases the number of discharges to 10.3 on average. In contrast, additional ES do not elicit the same responses.

ES patients are emergent by their nature and must be admitted, thus it is appropriate that the number of admitted ES patients does not change based on surgical occupancy. However, it is significant that the same early discharges do not seem to occur as the number of ES patient admissions increase. Therefore, clinicians are using the demand for SS patients as a lever to effectively increase capacity, but some of these changes are not benefiting the sicker ES patients, who are more likely to be negatively affected by delays in the start of care (Berry Jaeker and Tucker 2014, Chalfin et al. 2007). Furthermore, ES patients generally must wait in the emergency department (ED), thus increasing congestion in the already busy EDs (Falvo et al. 2007). These results suggest that there is a large benefit to predictability of incoming patients, and if we can increase the predictability of ES patients at all, there could potentially be significant benefits in care. We provide a discussion of possible solutions and their effects in the discussion of the paper.

This work contributes to the operations management literature in empirically quantifying the effects of prioritization and predictability among incoming patients. First, we identify that the type of patient admission moderates the clinician response to workload. These results suggest that it is not just the clinical condition or workload which affects clinician behavior, but also the more fundamental arrival type. Second, although SS patients provide a managerial lever to reduce demand when a hospital is nearly full, it is only employed at very high occupancy levels, while early discharges due to high SS demand occur at lower occupancy levels. Finally, we show that there exists a tension between the priority and predictability of these different incoming patient types. While the emergent patients are the most in need of a fast start to treatment, and medically should have priority, which is often assumed in queuing models, we provide empirical evidence that it is more likely that this on-time treatment occurs when a patient is more predictable. Moreover, these results show that predictability not only improves the ability to predict demand, but can also change the service level of other patients in the hospital.

4.2 Clinical patient care process

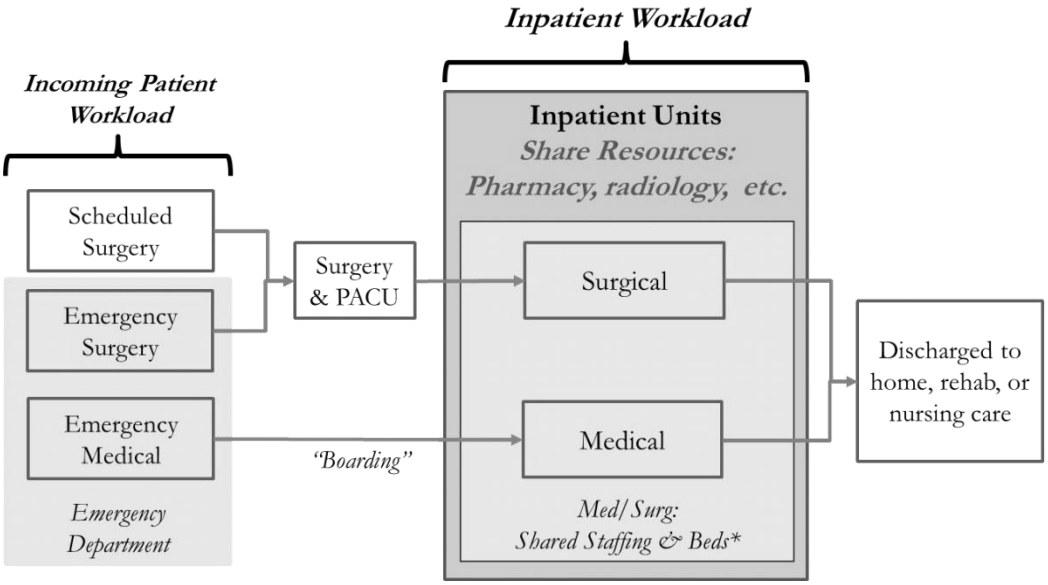
To understand why and how we categorize patients by type, it is critical to understand the care process of a patient through the hospital. All admitted patients arrive either through the emergency department (ED) or via scheduled admission (almost always surgical in nature). For a SS patient, the patient goes right to the pre-operative area at an

appointed time, has the surgery completed, awakens in the post anesthesia care unit (PACU) for 1-2 hours, and is then moved to the inpatient surgical unit. Meanwhile, an ES patient has a similar process, but first the patient goes to the ED where she is stabilized and determined to need surgery. Once stabilized, the patient is admitted to the surgical service in the hospital, and the care of the patient is now the responsibility of the surgical team. However, the patient may have to physically wait in the ED until the surgical team is ready to operate, or if the operation will not occur for a few hours, until a surgical inpatient bed is available (this process is known as "boarding"). Once the patient has the operation, the process is nearly the same, though there may be a slightly longer stay in the PACU if a bed is not ready. EM patients, like ES patients, are first seen in the ED where they are stabilized and determined to need admission to the hospital. Once stabilized, they are admitted to the medical service of the hospital. Like ES patients, EM patients often have to board in the ED until a bed is ready in an inpatient hospital unit.

Based on his or her diagnosis, a patient is assigned to a specific hospital service, which is the specialty responsible for caring for the patient. Three broad categories of patient services exist: medical, surgical, and obstetrics. Although some hospitals, particularly larger, academic centers, have subspecialties, these services are not standardized across hospitals, so we use the above broad clinical service categorizations. Each service has its own physicians, and the primary decisions regarding care are generally segregated by service (Gittell et al. 2000, Meyer 2011). However, it is common for lower acuity medical and surgical patients to be co-located on a general nursing unit, which can lead to a spillover of workload-related across patient types (Berry Jaeker and Tucker 2014) (See **Figure 4.2.1**).

In this study, we are most concerned with the how the number and type of incoming patients are affected by and affect currently convalescing patients. Therefore, we are interested in the workload of incoming and convalescing patients for each day in our study period. WE use the incoming occupancy as defined as the number of patients, by type, admitted to each hospital each day as the measure of the incoming patient workload. The number of admitted patients is calculated as both the absolute number as well as the percentage of maximum admitted for any day in that hospital (see Section 4.5 for more detail). Similarly, inpatient/convalescent patient workload is defined as the number of free beds or the percentage of beds occupied, by patient type, for each hospital each day.

Figure 4.2.1: Patient care process through the hospital



To categorize each patient as either medical or surgical, we rely on the Center for Medicare and Medicaid's (CMS) diagnostic related group (DRG) and major diagnostic category (MDC) classifications. Upon completion of a patient's hospital admission, each patient is assigned one of 746 DRGs based on the clinical diagnosis and primary needs of the patient, including the patient's severity level (Office of Inspector General 2001). While there is some variation among patients of a given DRG, CMS treats it as minimal, and these DRGs are used to calculate a payment level from CMS, where CMS reimburses for the treatment of the condition, irrespective of the costs incurred (Office of Inspector General 2001). Each DRG is a patient subtype within one of 25 MDCs, such as respiratory system or pregnancy, childbirth, and puerperium. In addition to the MDC, each DRG is broadly defined to be of type medical or surgical. Since obstetrics patients are treated separately from medical and surgical patients, they are not included in this analysis, so we remove them from the sample. Of the remaining patients, we classify each as surgical or medical based on his or her DRG.

4.3 RELATED LITERATURE

In this paper, the primary focus is on how the type of patient admission affects the probability of admission for these incoming patients, and the LOS of convalescing patients in hospitals, particularly under periods of high patient load. This work relates to the broader work on how to improve service quality and throughput in service settings.

At a high level, the role of workload in hospital care is analogous to understanding how demand affects these service times and quality. We are also incorporating how predictability interacts with these workload-related effects. Thus we draw on two streams of literature: workload effects in services, and how predictability affects capacity allocation, scheduling, and productivity.

Traditional operational models assume that the time to complete the work on a particular unit (or customer) is a random variable with a constant mean in which each unit of work is independent from each other unit (Dallery and Gershwin 1992). More recently, however, there has been a growing collection of research which shows that the time to complete a set of tasks, or service, is in fact dependent on the workload of the system (e.g., KC and Terwiesch 2012, Powell and Schultz 2004, Tan and Netessine 2014), and thus affects quality (Oliva and Sterman 2001). In general, most of this work has focused on service settings where workers have discretion over the tasks completed for each customer (Hopp et al. 2009), and in particular healthcare settings (Batt et al. 2012, KC and Terwiesch 2009, Kim et al. 2012, among others) where there is high variability in demand and high clinician discretion in care (Eddy 1984). These studies show that as a hospital gets busier patient care changes (Freeman et al. 2014). For example, Berry Jaeker and Tucker (2014) find that busyness at the start of a patient's stay leads to an increase in LOS, while at the end of a patient's stay, it results in a shorter LOS. These early discharges have been shown to result in poorer patient outcomes with increased probabilities of readmission (Anderson et al. 2012, KC and Terwiesch 2012). Moreover, even if the patient is not discharged early, the high workload results in less clinician time per patient, which has been shown to result in higher mortality rates (Kuntz et al. 2013) and lower reimbursement (Powell et al. 2012).

In addition to the load of patients in the hospital, other research focuses on how incoming patients (or more generally, customers) during periods of high demand further affect service quality and throughput. Analytically, it has been shown that shorter service time, and thus lower quality, is optimal when the costs of waiting are high (George and Harrison 2001, Ha 1998, Hopp et al. 2007, Stidham and Weber 1989). Empirically, patients are more likely to be discharged early when the workload in a hospital inpatient unit is high and there are new patient admissions (KC and Terwiesch 2012), and there is work that models a discharge policy which optimizes utilization given a unit's occupancy and admission rate, as well as the cost of readmissions due to early discharges (Chan et al. 2012). In addition, to early discharges, there is evidence that as the hospital

gets busier, the clinical threshold to admit a patient gets higher (KC and Terwiesch 2014, Kim et al. 2012). However, for all of the work on the effects of patient admissions, there has been little work into the patient characteristics (beyond clinical diagnosis) which could moderate the response to high inpatient and incoming patient load.

While there has been little work on how the types of incoming patients affects workload-related changes in care, there has been research on how to schedule and allocate resources given scheduled and emergent demand for surgeries. Most of this work builds off of the broader revenue management literature on capacity allocation given consumers with different priority classes and levels of certainty (see Netessine and Shumsky 2002 for a review). In general, the research focuses on models for scheduling operating room (OR) (Ferrand et al. 2010, Gerchak et al. 1996, Guerriero and Guido 2011, Gupta 2007), diagnostic equipment (Green et al. 2006, Patrick et al. 2008), inpatient bed (Best et al. 2013, Litvak 2010), or clinic (White et al. 2011) time to maximize utilization of these expensive resources given different patient priorities and/or certainty of demand. This research highlights the tension between scheduled patients who are easier to plan for and emergency patients who have higher prioritization. Additionally, even a small increase in the certainty of demand can increase productivity significantly (Fisher and Raman 1996). While these models have provided a foundation for optimizing utilization under this tension, they do not account for any downstream clinician response to this demand, such as early patient discharges, when developing optimal scheduling policies, as there has been little empirical or analytical research into these effects. In addition, there is poor communication and integration across units within hospitals, so work tends to focus on one area of a hospital (Green 2012). Given that these workload-related responses could have significant implications scheduling surgical patients, as well as utilization of expensive hospital beds, we are interested in identifying how patient characteristics affect inpatient care.

4.4 Hypotheses

4.4.1 Probability of scheduled surgical admission

Hospitals are often run at high occupancy levels due to the high fixed cost nature of hospital care (Roberts et al. 1999). These high costs have resulted in the reduction of hospital beds while demand increases (State of California OSHPD Health Information Decision 2008) to increase the average occupancy. While high utilization is an important

goal for efficiency, the high variability of arrivals and treatment times, as well as the high costs in delaying treatment, make it difficult to ensure appropriate waiting times as average occupancy increases (Green and Nguyen 2001). In fact, the previous research into workload-related effects on care has shown that one response to high wait times is earlier discharges (KC and Terwiesch 2009). However, there may be another managerial lever for managing excess demand: SS patients. By definition, SS patients are more predictable than emergency patients. In addition, they are also not emergent. Therefore, the cost of delaying surgery for such a patient is much less than for an emergency patient, where medical literature suggests that even small delays in care can result in higher mortality and longer LOSs (Chalfin et al. 2007). Operations management research has theorized that this may be occurring (Kuntz et al. 2013), but to our knowledge has never been shown. Medical research suggests that scheduled inpatient surgeries are cancelled around 10% of the time (Argo et al. 2009, Hand et al. 1990), and that some of these cancellations are due to the unavailability of ORs and beds. However, these studies do not separate availability of a PACU bed from an inpatient bed, nor the direct correlation between cancellation and hospital occupancy. We hypothesize that as the surgical census in the hospital approaches it maximum, fewer SS patients will be admitted. Formally,

HYPOTHESIS 1A (H1A): *As the number of convalescing surgical patients in a hospital approaches the number of surgical beds, fewer SS Patients will be admitted.*

In addition to the number of beds currently occupied in the hospital, there are also incoming emergency patients who will need beds. In many hospitals, the medical and surgical beds are actually shared across patient types, so that ES or EM patients can occupy any available medical/surgical bed. Both ES and EM patients must be prioritized over SS patients to receive care quickly. As a result, if the number of incoming patients is larger than the number of available beds, then SS patients will have to be cancelled.

HYPOTHESIS 1B (H1B): *As the number of emergent incoming patients in a hospital increases, fewer SS Patients will be admitted.*

4.4.2 EARLY DISCHARGE

When the number of incoming patients in combination with the currently convalescing patients exceeds the maximum capacity of the hospital, then there must be a reduction

in the number of patients needing a bed. As we described in Hypotheses 1a and 1b above, one lever is to reduce the number of incoming SS patients. However, cancelling a scheduled surgery is not without cost, since patients may have started to prepare, and could potentially be suffering from pain. In addition, delaying a surgery can result in patient abandonment and lost profits (Macario et al. 2001), so hospitals do not want injure goodwill by repeatedly cancelling SS patients. Yet, if there are more incoming patients than available beds, the only other option is to discharge a patient early. Previous research has shown within a busy ICU that as admissions increase, a patient is more likely to be discharged early to a less intensive inpatient unit (KC and Terwiesch 2012). Even if the unit is not completely full, and the available beds could accommodate all of the incoming patients, we could imagine that a high number of admissions could still result in early discharges. Patient admissions are time consuming tasks that require significant mental capacity (Berry Jaeker and Tucker 2014). Therefore, a high number of new patients would incentivize a clinician to discharge current patients to alleviate some of the workload and reduce potential queuing for the incoming patients (Oliva and Sterman 2001). Moreover, clinicians in an inpatient unit may want to ensure that they will have capacity for any additional emergency patients, and will proactively discharge patients near the completion of their stays. Thus, we hypothesize

HYPOTHESIS 2A (H2A): *As the number of SS or ES patients increases during a day, more convalescing surgical patients will be discharged.*

We predict that, in general, more surgical admissions should result in more surgical discharges. However, these incoming patients have different characteristics that could moderate the impact they have on early discharges. Specifically, SS patients are by definition predictable. Nurses and physicians know in the morning the number of scheduled surgeries expected each day. Consequently, they know if they will have a high workload during the day, and can thus plan accordingly. The discharge process for a patient can be sped up by clinicians: physicians can order a discharge at any time they feel the patient is ready to go home, and nurses can complete many discharge-related tasks (such as patient education) well before the patient actually leaves the hospital. Analytically, Armony and Gurvich (2010) have shown in call centers that knowing a customer will be coming in reduces the services that a worker should provide for a current customer. Therefore, we would expect for the number of discharges associated with additional SS patients to be greater than for an additional ES patient.

HYPOTHESIS 2B (H2B): *An additional SS patient has a greater effect on the number of discharged surgical patients than an additional ES patient.*

Discharging fewer patients as a result of incoming ES patients could mean that there are fewer readmissions. However, it could be coming at the cost of delays in treatment for incoming ES patients due to high workload, whereas SS admissions increase early discharges so that these patients are less affected by high workload. If there are delays in treatment, then we expect there to be longer LOSs for ES patients (Berry Jaeker and Tucker 2014, Chalfin et al. 2007). Thus, we hypothesize:

HYPOTHESIS 2C (H2C): *ES patients are more likely to have a delay in surgery due to high incoming workload than SS patients.*

4.5 DATA

We use patient-level records for all inpatient discharges in the state of California between December 2007 through December 2009 (Office of Statewide Health Planning and Development). We sum all patient admissions and subtract all patient discharges for all patients admitted in December, 2007 to determine the baseline number of patients in each hospital on January 1, 2008. We restrict our sample to only include patients admitted after December 31, 2007, and before November 30, 2009 since our data only has patients who were discharged by December 31, 2009, and does not include any patients admitted in December, but discharged on or after January 1, 2010, thus making it impossible for us to get an accurate census during that month. Out of the 448 hospitals in the data set, we limit our data to acute care hospitals with (1) at least 3500 patients over the course of our study (23 months) to ensure enough patients per day to calculate reasonable occupancy levels; (2) an average occupancy of obstetrics patients of at least 40% to guarantee more than intermittent obstetric patients; (3) a 24 hour ED to ensure there are emergency admissions; and (4) median daily (weekday) ES and SS admission rates of at least 3 patients each. This resulted in a sample of 169 hospitals. Each admission is its own record and includes the date admitted, date discharged, demographic information on the patient (e.g. gender, race, age), hospital of care, diagnoses (including comorbidities, which we categorize using the Elixhauser Index (Elixhauser et al. 1998)), major procedures, disposition (i.e., discharged home, to home health services, etc.), if an admission was scheduled or not, and DRG, among others. Consequently, we know whether the patient was medical, surgical, or obstetric, and if

the visit was scheduled or emergent. Lastly, we have dates for all major procedures performed, and the LOS of each patient in days. We measure LOS as day of discharge minus day of admission. However, because we have the day of each major procedure, for surgical patients we can also measure LOS as date of discharge minus date of surgery. We use the latter in as a robustness test to determine if delays in the start of care increase the LOS beyond the dely.

The first method we use is to segment the sample into similarly sized hospitals. Specifically, we place hospitals into 11 "bins" based on the median weekday surgical census, with each bin representing hospitals that have median weekday surgical censuses between 1-10, 11-20, 21-30, etc., with the last bin having all hospitals with a median weekday surgical census over 100. To ensure there are consistently incoming SS patients, we only use hospitals that have at least 31 patients. In addition, since the hospitals that have a median weekday surgical census over 100 are rarer and significantly different sizes, so we exclude them from our sample. Of the remaining hospitals, we want to ensure that the median weekday medical census is comparable across hospitals within a bin. Thus, we remove hospitals with median weekday medical censuses below the 5^{th} percentile and above the 75^{th} percentile (the distribution is skewed with a long right tail). While this method reduces the hospitals in our sample, it allows us to do more refined analysis where we are able to test the effect of each additional patient. Although the hospitals in each bin are similar in size, they are not exactly equal. To account for this in determining the effects of census on the numbers of admissions and discharges, as well as the LOS of patients, we use a method similar to KC and Terwiesch (2014). In particular, we define the surgical (medical) census as the number of available surgical (medical) beds, at the start of the day, where the number of available beds is the maximum census value minus the number of beds currently occupied, up to the 99^{th} percentile of available beds. We follow this method because there is still some difference in size across the hospitals, but the effect of 1 available bed should be relatively equal for each. Note, we do not adjust the number of incoming patients or discharged patients from the absolute number that are admitted or discharged over the course of a day.

The second method we use allows us to include the whole sample to strengthen our results that these effects occur across more than 200 hospitals. Since the hospitals in our data vary greatly in the number of admissions and number of beds, the absolute level of inpatient workload and incoming patient load can similarly vary. To be able to analyze the effects of patient workload across these heterogeneous hospitals, we use two

different methods. In the first we convert the absolute number of medical and surgical patients currently convalescing in the hospital (as the start of the day), as well as the number of incoming medical and surgical patients (throughout the day) into occupancy levels. These occupancies represent the percentage of the maximum for each patient workload category for that hospital. While we would prefer to calculate occupancy based on staffed beds available for a given day, the recorded numbers of staffed beds are at the year level, and often quite different from reality. Consequently, following the methods of (Berry Jaeker and Tucker 2014, Kuntz et al. 2013), we divide the number of visits by the maximum number of daily patient visits, and refer to this value as occupancy. Specifically, we calculate the 99th percentile of the number of patients treated (or admitted for incoming patients), by type, over the previous 90 days, based on whether it is a weekend or weekday. We use the 99th percentile to account for any extraordinary circumstances that do not actually reflect the number of beds generally available for occupancy. We separate weekday from weekend to account for the "weekend effect" where scheduled patients are nearly always admitted on weekdays. Thus, to calculate the occupancy for hospital "H" on Monday, June 9, 2008, we first find the 99th percentile of the number of patients who were inpatient during weekdays in hospital H from March 11, 2008 through June 9, 2008, and label that the maximum. We then divide the number of patients actually in the hospital on June 9 by that maximum.

For our analysis of the number of admissions and discharges, we use hospital-day level data. However, for the LOS, we need patient-day level data. We are interested in the effects of incoming and convalescing patients on surgical patients, so we focus our analysis on five surgical DRGs for a total of 186,500 patient visits (see **Figures 4.5.1 and 4.5.2** for Summary statistics). We restrict our analysis to these DRGs since they are among the 20 most common DRGs with average LOSs (as defined by CMS) of at least 3 days, but less than 8 days (See Figure 4.5.2 for DRGs used). This average LOS range ensures that the LOS is long enough to detect day level changes in LOS, but not so long that the inherent variability in LOS longer than 7 days makes it difficult to predict the expected LOS for patients, which reduces noise in our sample. By analyzing five DRGs, we ensure that our results are robust across clinical patient types. In the DRG classification, there are multiple DRG numbers for the same primary complaint, but they differ in severity levels. Each of the five DRGs in our study represents a specific primary reason for visit and intensity level selected because its frequency of occurrence and average LOS met the above described criteria for inclusion.

Figure 4.5.1. Summary statistics(*Hospitals=169)*

Median occupancies (%)	Average	Std dev	Minimum	Maximum
Surgical	75.7	6.2	48.1	89.7
Medical	78.5	5.4	52.2	100
SS admissions	51.4	8.3	17.6	72.7
ES admissions	44.4	8.6	0	76.9
EM admissions	61.9	7	40	96.4
Day of Week	% of Total			
Monday	24.3			
Tuesday	24.3			
Wednesday	19.5			
Thursday	17.7			
Friday	14.2			

To further reduce noise from our analysis of the impact of workload on a patients' LOS, we exclude patients who die during their stay since it is very difficult to predict the LOS of a patient who ends up dying in the hospital (Kuntz et al. 2013), and the death of a patient should be uncorrelated with the LOS of the remaining patients outside of the effect of occupancy and only 0.50% of the patients in our DRGs of interest died within the hospital in our sample. We also exclude patients who are transferred to another hospital, as these patients have LOSs that are impacted by factors that are not directly related to the occupancy of the hospital, as there is little evidence that transfers take place due to hospital occupancy (Kuntz et al. 2013). Transfers also represent less than 2% of the total patients within these ten DRGs. Moreover, we exclude patients who left against medical advice as it is difficult to predict when they will leave, and they are only 1.10% of the sample. Note that while we focus on five DRGs for calculating effects of workload on LOS, the inpatient and incoming patient volumes include all DRGs because the incoming work and current occupancy levels in the hospital are comprised of all patient types, including those who died, were transferred, or left against medical advice.

Figure 4.5.2. DRGs used and number of patients of each type

DRG	*Name*	*Number of Patient Visits in our sample*	*Average LOS among US Medicare Pts (Days)*
470	Major joint replacement or reattachment of lower extremity w/o MCC	119,025	3.8
460	Spinal fusion except cervical w/o MCC	23,067	4.0
481	Hip & femur procedures except major joint w/CC	12,596	5.7
419	Laparoscopic cholecystectomy w/o CDE w/o CC/MCC	19,656	3.0
742	Uterine & adnexa proc for non-malignancy w/CC/MCC	12,156	4.3
	TOTAL	186,500	

4.6 ECONOMETRIC SPECIFICATION

We are first interested in the number of SS admissions given then number of patients currently in the hospital, as well as the number of other admissions. The number of patients admitted or discharged can be thought of as a count function. Our data is not overly dispersed, so we use a Poisson regression to model the count function. In all models, we cluster the standard errors by hospital. Since SS admissions almost exclusively occur on weekdays, we limit our study to Monday through Friday arrivals. We therefore have,

$$Ln(SS_{d,b}) = \beta_1 Controls_d + \beta_2 Surg_d + \beta_3 (Surg_d)^2 + \beta_4 Med_d + \beta_5 (Med_d)^2 + \beta_6 ES_d + \beta_7 (ES_d)^2 + \beta_8 EM_d + \beta_9 (EM_d)^2 + \varepsilon_{d,b} \; if \; bin = b \quad (4.1)$$

$SS_{d,b}$ is the number of SS admissions on day, *d*, for a hospital in bin, *b*. *Controls* includes the factors, that in addition to our variables of interest, affect the number of admissions that we must take into account. Specifically, we want to control for the median number of SS admissions (*MedianAdmitSS)* for each hospital as some hospitals may simply have higher patient turnover. We create a categorical variable for MedianAdmitSS to account for any unobserved differences for differently sized hospitals. Also, we control for the day of the week (*DOW)* as the number of scheduled surgeries has been shown to be higher earlier in the week (Litvak 2010, McManus et al. 2003) since physicians want to be able to discharge their patients before the weekend. To control for trends and seasonality, we have a dummy variable for the quarter-year combination. Finally, to control for the number of patients who are near the beginning of their stays, and thus unlikely to be discharged, we include the number of patients admitted (*Admit*) the day before (d-1), as well as it's square. $Surg_d$ is the number of available surgical beds on day *d*. Med_d is the number of available medical beds on day. For these census variables, we would expect the coefficient to be positive (i.e. as there are more available beds, there are more SS admissions). ES_d and EM_d are the number of ES and EM admissions on day *d*, respectively. We predict that the coefficients on ES and EM will be negative (i.e., as the number of ES and EM admissions increase, the number of SS admissions decrease). For our variables of interest, as well as the number of previous admissions, we include the main effects as well as a square time to account for any non-linearity. Moreover, it will allow us to see if

there are tipping points at which the policy goes from early discharges to SS cancellations.

We perform a similar analysis with a Poisson regression using occupancies (percentages) to be able to compare across hospitals. Rather than use the median number of SS admissions, we include a hospital dummy to capture any fixed effects. In addition, we do not use the number of available Surgical or Medical beds, instead the *Surg* and *Med* variables represent occupancies, which we still square to account for any non-linear effects. Thus, we would expect the coefficients to be negative in this formulation (i.e., as medical and surgical occupancies increase, the SS admission rate will decrease). This leaves us with the following specification:

$$Ln(SS_{d,b}) = \beta_1 Controls_d + \beta_2 Surg_d + \beta_3 (Surg_d)^2 + \beta_4 Med_d + \beta_5 (Med_d)^2 + \beta_6 ES_d + \beta_7 (ES_d)^2 + \beta_8 EM_d + \beta_9 (EM_d)^2 + \beta_{10} Hosp + \varepsilon_{d,b} \quad (4.2)$$

For both of these models, we run a robustness where we replace ES to be the dependent variables, replace median SS admits with median ES admits, and include SS as an independent variable. We want to confirm that the number of emergency surgeries is not affected by SS admissions.

Next we're interested in the number of surgical discharges based on the busyness of the hospital and the number of incoming patients. We use a similar method as above, but for the Poisson models, we include a measure of the absolute number of surgical patients ($AbsSurg_d$) currently in the hospital. We do this because this number limits the number of possible discharges on a given day. In addition, instead of the median number of SS admissions, we include the median number of discharges. Since the number of discharges is generally greater than 0, we include a constant term in this model. Finally, we also add a term for the number of SS admissions, as this is one of our primary variables of interest. Thus, we have

$$Ln(SurgDC_{d,b}) = \beta_0 + \beta_1 Controls_d + \beta_2 Surg_d + \beta_3 (Surg_d)^2 + \beta_4 Med_d + \beta_5 (Med_d)^2 + \beta_6 ES_d + \beta_7 (ES_d)^2 + \beta_8 EM_d + \beta_9 (EM_d)^2 + \beta_{10} SS_d + \beta_{11} (SS_d)^2 + \beta_{12} AbsSurg_d + \varepsilon_{d,b} \quad (4.3)$$

$if\ bin = b,$

where $SurgDC_{d,b}$ is the number of surgical discharges on day d for a hospital in bin b.

Again, we run another Poisson regression using occupancy measures, but due to the difference in occupancy definitions, we do not include the $AbsSurg_d$ term, and instead of the median number of discharges, we include a hospital fixed effect. Therefore, we are left with

$$\begin{aligned} Ln(SurgDC_d) &= \beta_1 Controls_d + \beta_2 Surg_d + \beta_3 SurgBusy_d \\ &+ \beta_4 Med_d + \beta_5 MedBusy_d + \beta_6 ES_d + \beta_7 ESBusy_d + \beta_8 EM_d \\ &+ \beta_9 EMBusy_d + \beta_{10} SS_d + \beta_{11} SSBusy_d + \beta_{12} Hosp + \varepsilon_{d,b} \end{aligned} \quad (4.4)$$

Finally, we are interested in understanding the effect of these early discharges on incoming surgical patients' time to procedure, based on admission type, which is a good measure of the start of care. For this analysis, we restrict our sample to incoming patients of the five DRGs described above, and to patients who have a LOS of at least 3 days. Unlike the previous cases, only occupancy (percentage) measures are used due to sample size limitations. Unlike in the previous, this model employs a probit regression where the dependent variable is the primary procedure being performed on the day of admission (outcome=0) or on a later day (outcome=1). In addition, we are interested in the effect admission type (either scheduled or emergent) has on the propensity for a surgery to be delayed, so it is included as a control. For hypothesis 2c, the variable of most interest are the effects of ES and SS admissions on the probability of delayed surgery. Since the probability of the day of surgery is also likely affected by admission type of the patient (emergency or scheduled), we run the same model, but interacting the variables of interest by admission type. Each of the DRGs is its own model, and is as follows

$$\begin{aligned} \Pr(Y = 1)_i &= \beta_1 Controls_i + \beta_2 Surg_d + \beta_3 (Surg_d)^2 \\ &+ \beta_4 Med_d + \beta_5 (Med_d)^2 + [\beta_6 ES_d + \beta_7 (ES_d)^2] * Emergency_i \\ &+ \beta_8 EM_d + \beta_9 (EM_d)^2 + [\beta_{10} SS_d + \beta_{11} (SS_d)^2] * Emergency_i \\ &+ \beta_{12} Emergency_i + \beta_{13} Hosp + \varepsilon_i \end{aligned} \quad (4.5)$$

4.7 RESULTS

Modeling the number of SS admissions using the census levels, we find support for H1a that as the number of convalescing surgical patients approaches the number of surgical beds, the number of SS admissions declines, but to a lesser extent as the number of available beds increases. For ease of interpretation, we only show the results for hospitals in surgical bin 4, one of the most common surgical inpatient unit sizes. In

Figure 4.7.1, the main effect of available surgical beds on SS admissions is 0.0965 ($p<0.01$), while the coefficient for the square term is -0.0044 ($p<0.01$). Thus, for a hospital that has 5 expected SS admissions if the number of available surgical beds goes from 5 to 4, an additional surgical patient reduces the expected number of SS admissions by 0.28 to 4.72. Similar results are found when using occupancy measures across all hospitals.

Figure 4.7.1. Poisson regression of number of SS admissions (census levels)

Ln(number of SS admissions)	**Coefficient**
Available Surgical Beds	0.0965**
	(0.011)
(Available Surgical Beds)^2	-0.0044**
	(0.001)
Available Medical Beds	0.0634**
	(0.011)
(Available Medical Beds)^2	-0.0028**
	(0.001)
ES Admissions	0.0252+
	(0.015)
(ES Admissions)^2	-0.0051*
	(0.002)
EM Admissions	0.0314**
	(0.005)
(EM Admissions)^2	-0.0008**
	(0.000)
Controls	
Quarter-Year	Included
Day of Week (Baseline: Monday)	
Tuesday	0.0612
Wednesday	-0.0828**
Thursday	-0.1851**
Friday	-0.3373**
Median SS admissions (Baseline: 3)	
4	0.4641**
5	0.5456**
6	0.7172**
7	0.8424**
8	0.9636**
9	1.1416**
Admissions previous day	-0.0003
(Admissions previous day)^2	0.0005
n=	5,270

Median surgical census is 31-40; Robust standard errors in ( ); ** $p<0.01$, * $p<0.05$, + $p<0.1$

Reducing the number of available surgical beds reduces the number of expected SS admissions whenever there are fewer than 11 available surgical beds. Above that, the number of SS admissions is unaffected or greater. **Figure 4.7.2** shows that the coefficient for the main effect of inpatient surgical occupancy is 0.0145 ($p<0.01$) while the coefficient for the square term is -0.0001 ($p<0.01$). Thus, the effect of a 5% increase in surgical

occupancy from 90% to 95%, given an initial expected number of SS admissions of 5 is to reduce the expected number of SS admissions by 0.1 patients. For models 1 and 2, we also run a zero-inflated Poisson regression (zero-inflation of day of week and quarter of year), and the results are consistent.

Figure 4.7.2. Poisson regression of number of SS admissions (occupancy levels)

Ln(number of SS admissions)	**Coefficient**
Surgical occupancy	0.0145**
	(0.002)
(Surgical occupancy)^2	-0.0001**
	(0.000)
Medical occupancy	0.0035
	(0.003)
(Medical occupancy)^2	-0.0000
	(0.000)
ES occupancy	-0.0024**
	(0.001)
(ES occupancy)^2	0.0000*
	(0.000)
EM occupancy	0.0038**
	(0.001)
(EM occupancy)^2	-0.0000+
	(0.000)
Controls	
Hospital	Included
Quarter-Year	Included
Day of Week (Baseline: Monday)	
Tuesday	0.2192**
Wednesday	0.1130**
Thursday	-0.0058
Friday	-0.1760**
Admissions previous day occupancy	0.0164**
(Admissions previous day occupancy)^2	-0.0001**
n=	27,082

Robust standard errors in ( ); ** p<0.01, * p<0.05, + p<0.1

The results also provide support for H1b that as the number of EM and ES admissions gets high, the number of SS admissions decrease. The main effect coefficient for ES admissions in the census model (**Figure 4.7.1**) is 0.025 (p<0.1), while the coefficient for the square term is -0.0051 (p<0.05). Therefore, with an expected number of SS admissions of 5, if the number of ES admissions increases from 9 to 10, the expected number of SS patients falls by 0.35 patients to 4.65. The tipping point where the number of SS patients begins to decrease, given 5 expected SS admissions is between 2 and 3 additional ES patients. Similar results are found for EM patients, with coefficients on the main and square effects of 0.0314 (p<0.01) and -0.0008 (p<0.01), respectively. However, in this situation, the tipping point is at between 20 and 21 EM admissions, which means

that EM admissions are unlikely to reduce the number of SS admissions. It is comforting to note (and important for our model) that the reverse does not hold: the number of SS admissions does not affect the number of ES admissions. **Figure 4.7.2,** which shows the occupancy models, shows a statistically significant decrease in the number of SS patients admitted as the number of ES admissions increases (main effect coefficient of -0.0024, p<0.01), in support of H1b. However, the occupancy results do not support that increased EM admissions decrease SS admissions, which we will discuss in Section 8 below.

While the number of SS admissions are affected by heavy workload, this may not be the only mechanism for managing capacity. Hypothesis 2a says that increases in the number of SS or ES patients will result in additional surgical discharges. In the census model (**Figure 4.7.3**), there is partial support for the H2a with the number of SS patient admissions having a statistically significant effect on the number of surgical discharges (main effect coefficient of 0.0458, p<0.01; square coefficient of -0.0014, p<0.1). This effect

Figure 4.7.3. Poisson regression of number of surgical discharges (census levels)

Ln(number of SS admissions)	**Coefficient**
Available Surgical Beds	0.0291**
	(0.003)
(Available Surgical Beds)^2	0.0118
	(0.008)
Available Medical Beds	-0.0007
	(0.001)
(Available Medical Beds)^2	0.0208**
	(0.007)
ES Admissions	-0.0011**
	(0.000)
(ES Admissions)^2	0.0458**
	(0.013)
EM Admissions	-0.0014+
	(0.001)
(EM Admissions)^2	-0.0019
	(0.009)
Controls	
Quarter-Year	Included
Median Surgical Census	Included
Day of Week	Included
Admissions previous day	0.0028
(Admissions previous day)^2	-0.0001
n=	5,270

Median surgical census is 31-40; Robust standard errors in ( ); ** p<0.01, * p<0.05, + p<0.1

is equivalent to an increase of the expected number of patients discharged from 10 to 10.31 for an increase in SS admissions from 5 to 6. The tipping point where additional SS patient causes fewer surgical discharges, given the expected number of discharges is 10 is

16 to 17 additional SS patients, which will almost never happen in this sample where the hospital with the largest median number of SS admissions has a median of 9 SS admissions. There is no statistical support for ES patients in the census model, however, there is statistical support for both SS and ES admissions in the occupancy model (**Figure 4.7.4**). The coefficients of the effects of increased SS and ES occupancy on surgical patient discharges are 0.001 ($p<0.05$) and 0.0008 ($p<0.01$), respectively. Thus, an increase in the SS admission rate from 90% to 95% increases the number of surgical discharges from 10 to 10.1, while the same increase in ES occupancy increases the number of discharges from 10 to 10.08. Thus there is moderate to significant support that increased SS and ES admissions increase the number of surgical discharges. Moreover, the effect of increased SS admissions is greater than the effect of ES admissions on surgical discharges, in support of H2b.

Figure 4.7.4. Poisson regression of number of surgical discharges (occupancy levels)

Ln(number of SS admissions)	**Coefficient**
Surgical occupancy	0.0254**
	(0.002)
(Surgical occupancy)^2	-0.0001**
	(0.000)
Medical occupancy	-0.0055*
	(0.003)
(Medical occupancy)^2	0.0000
	(0.000)
ES occupancy	0.0030**
	(0.000)
(ES occupancy)^2	-0.0000**
	(0.000)
EM occupancy	0.0008*
	(0.000)
(EM occupancy)^2	-0.0000+
	(0.000)
Controls	
Hospital	Included
Quarter-Year	Included
Day of Week	Included
Admissions previous day occupancy	0.0079**
(Admissions previous day occupancy)^2	-0.0000**
n=	27,082

Robust standard errors in ( ); ** $p<0.01$, * $p<0.05$, + $p<0.1$

Figure 4.7.5 shows the effects of inpatient and incoming patient occupancy on delays in the start of treatment across DRGs, with (even numbered columns) and without (odd numbered columns) an interaction for admissions type. The variables of most interest are the effects of SS admissions and ES admissions, both interacted and not

Figure 4.7.5. Probit model of the probability that the day of surgery is after the day of admission

DRG	*470*		*460*		*481*		*419*		*742*	
Clinical Condition	*Major Joint Replacement or Reattachment of Lower Extremity w/o MCC*		*Spinal Fusion except Cervical w/o MCC*		*Hip & Femur Procedures except Major Joint w/ CC*		*Laparoscopic Cholecystectomy w/o CDE w/o CC/MCC*		*Uterine & Adnexa Proc for non-malignancy w/CC/MCC*	
Probability (Surgery day>0)	***(1)***	***(2)***	***(3)***	***(4)***	***(5)***	***(6)***	***(7)***	***(8)***	***(9)***	***(10)***
Surgical occupancy	-0.0112**	-0.0089**	-0.0161	-0.0135	-0.0150*	-0.0113	-0.0059	-0.0037	-0.0078	-0.0071
	(0.002)	(0.002)	(0.011)	(0.012)	(0.007)	(0.007)	(0.005)	(0.005)	(0.011)	(0.011)
(Surgical occupancy)^2	0.0001**	0.0001**	0.0001	0.0001	0.0001*	0.0001+	0.0000	0.0000	0.0000	0.0000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Medical occupancy	-0.0286**	-0.0201**	-0.0328**	-0.0283*	-0.0050	0.0090	-0.0097	0.0002	-0.0289**	-0.0280*
	(0.004)	(0.004)	(0.012)	(0.012)	(0.008)	(0.009)	(0.007)	(0.008)	(0.011)	(0.011)
(Medical occupancy)^2	0.0002**	0.0001**	0.0002*	0.0001+	0.0000	-0.0001	0.0001	-0.0000	0.0002**	0.0002**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
SS admission occupancy	0.0003	-0.0161**	0.0021	-0.0116+	-0.0016	-0.0326**	0.0005	-0.0179**	0.0039	0.0006
	(0.001)	(0.002)	(0.004)	(0.006)	(0.002)	(0.010)	(0.001)	(0.006)	(0.003)	(0.007)
(SS admission occupancy)^2	0.0000	0.0001**	-0.0000	0.0001+	0.0000+	0.0003**	-0.0000	0.0001*	-0.0000	0.0000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
SS admission occupancy*I.Emergency		0.0166**		0.0199*		0.0314**		0.0189**		0.0056
		(0.003)		(0.009)		(0.010)		(0.006)		(0.008)
(SS admission occupancy)^2 *I.Emergency		-0.0001**		-0.0001+		-0.0002**		-0.0001*		-0.0001
		(0.000)		(0.000)		(0.000)		(0.000)		(0.000)
ES admission occupancy	-0.0023*	-0.0017	-0.0053+	-0.0050	0.0018	0.0049	-0.0026	-0.0065	0.0070+	0.0058
	(0.001)	(0.002)	(0.003)	(0.003)	(0.003)	(0.007)	(0.002)	(0.005)	(0.004)	(0.005)
(ES admission occupancy)^2	0.0000	0.0000	0.0000	0.0000	-0.0000	-0.0000	0.0000	0.0000	-0.0001*	-0.0001
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
ES admission occupancy*I.Emergency		0.0004		0.0065		-0.0027		0.0059		0.0041
		(0.002)		(0.010)		(0.008)		(0.006)		(0.008)
(ES admission occupancy)^2 *I.Emergency		-0.0000		-0.0000		0.0000		-0.0000		-0.0000
		(0.000)		(0.000)		(0.000)		(0.000)		(0.000)
EM admission occupancy	-0.0033*	-0.0024	-0.0034	-0.0029	-0.0027	-0.0009	-0.0022	-0.0012	-0.0177**	-0.0173*
	(0.002)	(0.002)	(0.006)	(0.006)	(0.004)	(0.004)	(0.004)	(0.004)	(0.007)	(0.007)
(EM admission occupancy)^2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001*	0.0001*
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Controls										
I.Emergency admission	2.4760**	1.8670**	2.2364**	1.2908**	1.4646**	0.5480+	1.5451**	0.8457**	2.0667**	1.8645**
Hospital						Included				
Quarter-year						Included				
Day of Week						Included				
Previous day admissions						Included				
(Previous day admissions)^2						Included				
n=	208,693	208,693	22,357	22,357	12,376	12,376	19,466	19,466	11,794	11,794

Robust standard errors in (); ** p<0.01, * p<0.05, + p<0.1

with the admission type, on the propensity for a surgery to be delayed until the next day or later. When the admission type is not interacted with ES and SS admissions, there is little support for hypothesis 2c. However, once the interactions are included, hypothesis 2c is supported. An increase in SS admissions reduces the propensity of a surgery being delayed for scheduled patients ($p<0.01$ for three of five DRGs, $p<0.1$ for one DRG), while the same increase in SS admission increases the propensity of a delayed surgery for emergency patients ($p<0.01$ for four of five DRGs). In contrast, an increase in ES admissions does not change the propensity of surgery being delayed for any DRG or patient type. Thus, incoming ES patients does not affect either patient type, but ES patients are more likely to be delayed by increases in the number of incoming SS patients than SS patients are. In fact, higher numbers of SS patients seem to reduce the probability of delay of SS patients. We discuss the implications of these results in the next section.

4.8 Discussion and conclusion

The results of our study show that the admission type of a patient (i.e., emergent or scheduled), significantly affects the probability of own and other type admissions, and if admitted, the LOS of other incoming patients as well as currently convalescing inpatients. Specifically, we find that as the number of available surgical beds approaches zero, SS patients are more likely to be canceled, compared to the expected number of SS admissions, as well as compared to ES patients. Similarly, as the number of emergency admissions, either surgical or medical, increases, the number of SS admissions decreases. Given that SS patients are less urgent than ES, this practice is a method of managing hospital bed capacity constraints. However, we find it is not the only method for managing bed capacity. Prior to SS cancellations, there is evidence that the hospital first starts to increase discharges of currently convalescing surgical patients. Interestingly, hospitals are more likely to discharge patients early for SS admissions than for ES admissions. Relatedly, high numbers of SS admissions reduce the probability of delays in the start of surgery for scheduled patients, while the same increase in SS admissions increases the probability of a delay in surgery for emergency patients. These results suggest that the predictability of incoming SS patients allows hospitals to better prepare for their arrivals. Consequently, these less emergent patients are the ones who are actually more likely to receive timely, non-delayed, care, in contrast to the expected prioritization based on clinical condition.

The results of our study have significant managerial implications. In particular, we find that knowing that a patient will be arriving to an inpatient unit (i.e., a SS patient), incentivizes inpatient clinicians to discharge currently convalescing patients more quickly. Yet, it is the ES patients who need the prioritization of care. To address this issue, hospitals could have designated ORs and rooms for emergency patients to ensure there is enough capacity for ES patients when they need it, but this method

incurs the negative consequences of pooling, though there are strategies for dividing up hospital beds that take into account patient prioritization and can minimize the negative effects of pooling (Best et al. 2013).

Another solution is for hospitals to smooth out their inflow of schedule surgical patients. There is significant research in this area (Helm et al. 2011, Litvak 2010, McManus et al. 2003). While emergency admissions cannot be controlled, SS admissions can. Many studies have found that surgeries are scheduled non-optimally to maintain a relatively consistent inpatient occupancy level. Instead, surgeries are scheduled at the beginning of the week to ensure a patient is discharged prior to the weekend. Our own data is consistent with these results. However, by scheduling surgeries evenly throughout the week, not only would it reduce the number of inpatient and OR beds, it would reduce the number of cancelled surgeries, and could potentially lower the probability of an ES patient's care being delayed.

Another method for reaping some of the benefits of the predictability of SS patients is to start the admission process as soon as an ES patient enters the ED. Studies have shown that triage nurses are actually 80% accurate in predicting, at triage, if a patient will need to be admitted to the hospital (Saghafian et al. 2012). Therefore, it seems reasonable that the ED can give a fairly accurate estimate of the number of patients to be expected in surgery, and later in the hospital.

As with any study, ours has limitations. As mentioned in Section 4.7 above, we only find partial support that SS patients are cancelled due to a high number of incoming EM patients our census model. We theorize that this is because cancellations only occur when the unit is nearly full, or will be nearly full given the incoming patients (hence the large "tipping point"), which is at different occupancy levels for different hospitals. Our census models analyze very similarly sized hospitals, so our measure of available beds is more precise. In contrast, the occupancy models compare across more than 150 hospitals, and an occupancy of 95% for a small hospital may mean only one surgical bed is available, in a large hospital it could mean 5 beds are available, and thus a surgery is less likely to be cancelled. In general, we wish we could know exactly how many staffed beds are available at each hospital at the time of each patient's arrival. However, our data is limited to day-level censuses. Thus, we have tried to use both a modified census and occupancy levels to confirm our results.

In this study, we do not explore the effects of the early discharges on readmission rates. It is possible that some of these patients will be readmitted to the hospital. What is an appropriate readmission level, given the bed-hours saved is dependent on the LOS for a readmission, as well as the relative benefits of starting a patient's care earlier. Interestingly, the patients who suffer the least from delays in care, SS patients, are the most likely cause these early discharges. The greater benefit, which is more likely to offset the costs of readmissions, would come from early care of the incoming emergency patients. In addition, the emergency patients often must wait in the already crowded EDs and/or PACUs, so early

admission to surgery and/or the inpatient units would have spillover benefits to other units within the hospital.

Overall, this work provides significant insight to the response within hospitals for different types of incoming patients. There has been only limited research on how the characteristics of a patient affect the quality of care of other patients in a hospital. We show that the role of predictability goes beyond its effects on scheduling and capacity allocations, and results in behavioral responses to speed-up service times. However, the move to discharge patients early does not correspond to the highest priority incoming patients, suggesting that improving the "predictability" of emergency patients could have significant positive consequences in their quality of care, and more efficiently manage limited hospital capacity.

J. Berry Jaeker, Tucker, A.L., and M.H. Lee, 2013. Increased Speed Equals Increased Wait: The Impact of a Reduction in Emergency Department Ultrasound Order Processing Time. Working Paper. Harvard Business School. Boston, MA.

3

Increased Speed Equals Increased Wait: The Impact of a Reduction in Emergency Department Ultrasound Order Processing Time

3.1 INTRODUCTION

HEALTHCARE costs in the U.S. have skyrocketed since the 1980's. By 2010, healthcare expenditures accounted for 18% of the U.S. gross domestic product (Berwick and Hackbarth 2012). These high costs hinder industry competitiveness and drain financial resources away from other areas, such as education and transportation (Berwick and Hackbarth 2012, Fuchs 2012, Hussey et al. 2009). One of the largest opportunities for reducing healthcare costs is eliminating the use of medications, tests, and treatments that do not improve patients' health (Berwick and Hackbarth 2012, Gawande 2009, Hussey et al. 2009). Therefore, understanding factors that influence physicians' decision making can provide important levers for controlling costs. To date, misaligned financial incentives have been the primary explanation for why physicians order low-efficacy treatments. Physicians have discretion over the medical interventions prescribed for patients, which can result in overuse if they can increase their revenue by ordering more treatments (Gawande 2009, Hussey et al. 2009, Levin and Rao 2008). However, many physician groups, such as the one in our study, are not paid for each service provided and therefore have no financial

incentives to order additional tests, and yet low-efficacy treatments still occur in those settings. This suggests that other factors influence the use of these treatments.

Theory in the behavioral operations management literature provides an alternate explanation for why physicians order low-efficacy treatments: in a customer service environment where employees have high levels of discretionary task completion (DTC), service providers fill their time with work (Debo et al. 2008, Hasija et al. 2010, Parkinson 1958), and respond to increased capacity by *increasing* the number of services provided to their customers due to a reduction in the service's marginal cost (Hopp et al. 2007). Applied to a hospital environment, this theory predicts that if there is a process change that makes it quicker for physicians to order a medical test—which is equivalent to increasing the physician's effective capacity for ordering a test—physicians will respond by ordering the test for more patients. The dynamic exists even in the absence of direct financial incentives because providing more services to their patients is perceived as providing higher quality service. Thus, operations management theory suggests that process improvements that reduce the time required for physicians to order a test can result in higher use of tests, even those that are low-efficacy.

We build off of this theoretical work by exploiting an exogenous process change that occurred at one of two emergency departments (EDs) within the same health system and staffed by the same physician group. The ordering process for an ultrasound (U/S), a diagnostic test, was changed for the night and weekend shifts at one of the EDs so that it was less time-consuming for a physician to order an U/S than it had been previously. Therefore, in terms of the theory, because the change reduced the time it took a physician to order an U/S, it effectively increased physicians' capacity for ordering U/S. Over a three-year period, we studied ED patients with abdominal pain, a common symptom for which an U/S is one of several diagnostic options available to ED physicians. We find that the process change reduced the LOS of an ED patient who received an U/S by 21%, or approximately 1 hour. However, the change was also associated with a 70% higher probability of receiving an U/S. The net result was an increase in ED LOS across all patients. This was due to two effects: First an U/S is a lengthy study (patients who received an U/S had a LOS that was 30% higher than patients who did not). Second, even patients who did not receive an U/S had an increase in the time for other radiographic studies to be interpreted. As a result, the overall ED LOS for all abdominal pain patients increased. The longer LOS was associated with longer ED wait times to see a provider. Thus, it appears that the overall negative impact on patient flow eroded the benefit of shortening the ordering time. In addition, we found no evidence that the increase in U/S improved two standard measures of ED quality of care: admission rate to the hospital and readmission rate to the ED within three days. The only clinical change was a small (approximately 0.25 per patient) reduction in the number of laboratory tests ordered for patients after the processing change. Therefore,

the reduction in U/S ordering time resulted in a decline in the overall system performance, with a lower throughput rate, but no observed improvement in clinical quality.

Our work makes three contributions to the behavioral operations literature. First, we empirically validate that reducing the time it took physicians to order an U/S resulted in a higher probability of a patient receiving an U/S than in the past. This is consistent with Hopp, Iravani and Yuen's (2007) theory that when employees have discretion, reducing the marginal cost of providing a service will result in them giving that service to more customers. Our empirical validation of supply-induced demand highlights the importance of accounting for an increase in demand for a service, which is often thought of as exogenous, when increasing effective capacity of workers in DTC settings. Second, our study extends research on DTC settings to include the effects of changes in workers' capacity in a setting where resources are shared. We find that when increased use of the faster process placed additional demand on a shared resource (radiology), this slowed the care for other patients who used the same shared resource. These results suggest that a process improvement can inadvertently cause an increase in demand for a service as well as associated shared resources, which results in congestion, counter intuitively decreasing overall system performance. Our study thus highlights an important lesson for process improvement: reducing the cycle time of one step in a process can end up overloading a bottleneck resource (in our setting, radiology). Third, we examine the effects of changes in worker capacity on the system when incentives of individuals may not be aligned with the organization's goals. We show that while individual patients and physicians may benefit from the reduced processing time, there can be unintended consequences for overall system performance. These results illustrate how time-saving modifications to hospital processes—if they contribute to increased use—can paradoxically reduce productivity and contribute to rising costs.

3.2 LITERATURE ON THE IMPACT OF WORKER DISCRETION ON OPERATIONAL PERFORMANCE

There is a growing amount of research analytically and empirically modeling the effects of human behaviors on quality and productivity. This body of work addressed the call for research to expand our understanding of how workers behave, and how these behaviors affect the processes in which they work, particularly in service settings where workers have high levels of discretion over their tasks (Boudreau et al. 2003, Gino and Pisano 2007, Parkinson 1958). Prior studies have shown that processes are affected by workers' behavior, suggesting that behavioral effects must be accounted for when designing service processes.

Much of the analytical research in service settings has centered on the trade-off between service quality and processing time. This body of work is directly related to our study of physician ordering behaviors as physicians balance the care of multiple concomitant patients. When a service worker performs more tasks for his or her customer, it results in a higher quality experience for that customer

(Anand et al. 2011, Hopp et al. 2007). However, this higher level of service takes more time, which means that incoming customers will wait longer for service, decreasing the quality of their experiences. Thus, service quality can be increased, but at the expense of a longer processing time (Anand et al. 2011). Analytical models have shown that to optimize performance, the service time should remain the same or decrease for each customer as the number of customers in a system rise (George and Harrison 2001, Stidham and Weber 1989). This recommendation to reduce service time maximizes the utility for all customers. However, if given the choice, each individual customer would prefer a longer service time for himself, highlighting the tension that service workers face between maximizing the satisfaction for their immediate customer versus maximizing the average satisfaction of all customers (Ha 1998). In our research setting, this work is analogous to ED physicians trying to balance providing comprehensive services (i.e. diagnostic tests, procedures, symptom relief) that patients want while also being able to quickly get to the patients who are waiting to be seen.

Empirically, studies have validated the trade-off between service quality and speed. For example, Oliva and Sterman (2001) found that when congestion increased, back-office bank workers "cut corners" by spending less time processing loan applications, which resulted in lower quality evaluations and fewer approved loans, and consequently reduced the company's revenue. Other studies have shown that "speeding up" behavior is influenced by the visibility of the congestion (Schultz et al. 1999, Schultz et al. 1998, Song 2013), highlighting the sensitivity of human behavior to the state of the operating system. Much of the empirical work has looked at healthcare settings, perhaps because of the high levels of worker discretion and variability in treatment options (Eddy 1984). KC and Terwiesch showed that under periods of high patient load, employees worked faster (2009) and/or discharged patients earlier in order to free up bed capacity (2012). Discharging patients early is a form of reduced service quality as it has been shown to lead to readmission and rework (KC and Terwiesch 2012). Even if a patient is not discharged early, quality of care is negatively impacted by load: fewer tests are ordered (Batt et al. 2012); patients have longer LOS (Berry Jaeker and Tucker 2014); and worse clinical outcomes (Kim et al. 2012, Kuntz et al. 2013).

Understanding the impact of workload on employees' behaviors enables operations managers to allocate resources so they can better meet customers' needs. For example, for hospitals, information about the impact of workload on performance can be used to determine the desired occupancy levels, bed allocation decisions and staffing levels (Chan et al. 2012). More generally, Hopp, Iravani and Yuen (2007) examined the effect of an increase in the effective capacity of customer service providers with high levels of task discretion on waiting times. An increase in capacity could stem from process improvements such as increased training, new equipment that reduces processing time, or increased staffing. They found that increasing service capacity lowers the marginal cost to employees of providing the service, which in turn

motivates workers to provide these services to their customers. Receiving more services during their transactions increases customers' average processing times, which increases waiting times for customers queuing for service. More simply, lowering average processing times can paradoxically increase congestion in the system (Hopp et al. 2007). Further research has taken a more prescriptive approach and examined the impact of customer demand on the number of workers that should be hired and how their behavior should change in response to changes in customer demand. For example, Armony and Gurvich (2010) developed a model of call centers that provides a recommendation of when to upsell customers versus when to provide the minimum service level. In healthcare, there are models to predict the number of hospital beds needed to meet targeted service levels and waiting times (Green and Nguyen 2001), as well as research which highlights the negative consequences of delays in care (Chalfin et al. 2007) when service levels are not met. Collectively, these studies highlight the relationship between capacity and demand in a hospital and employees' behavioral response to any mismatches, as well as how this information can be used to allocate resources effectively. The studies suggest that employees will respond to available capacity by providing additional services, a theory which provides an alternate explanation for the use of low efficacy tests in healthcare: physicians might use the additional tests or treatments to improve their patients' perception of their quality of service. Our paper empirically tests these theories, as well as how changes in resource use affects shared resources in a hospital system.

3.3 STUDY SETTING

We exploit an exogenous process change that impacted only one of two partner EDs, a setting where employees have a high level of discretion over the level of service they provide to patients. The process change in the ordering of U/S by ED physicians at one, but not the other ED enables us to test the impact of a reduction in processing time in the use of a resource. Specifically, we use patient-level data from the EDs of two east coast academic hospitals, which we refer to as Flagship Hospital (Flagship) and Community Hospital (Community). Flagship is the largest hospital in the state and a Level 1 Trauma Center, and Community is a community hospital located five miles away. Both EDs are part of the same healthcare system staffed by the same physician practice group, and have a large patient load, with Flagship and Community having more than 100,000 and 60,000 emergency visits per year, respectively. The physician practice group is independent and is not employed by either hospital.

Over the course of our study period (September 30, 2009 – May 31, 2012), there were 83 attending physicians (also referred to as attendings) in the physician group, with 50 attendings practicing at both hospitals. Ultimate responsibility for all testing and treatment decisions was held by the attending physicians. In some instances, resident physicians (who are physicians undergoing additional post-

graduate training) or physician assistants made assessments of the patients and suggested possible treatments to the attending physician.

We focused our study on patients presenting to the ED with abdominal pain. We do so for several reasons. First, it is the most common chief complaint for patients who come to the ED (Pitts et al. 2008). Common diagnoses of abdominal pain include ulcers of the stomach, esophagus or small intestine; appendicitis; disorders of the gallbladder; kidney stones; inflammation of the colon; stomach or intestinal infection and constipation (Bengiamin et al. 2009). There are also rarer but life-threatening conditions that the ED physician must rule-out, such as a ruptured aneurysm of the abdominal aorta, an intestinal obstruction, a perforated stomach ulcer, inflammation of the pancreas and occlusion of the intestinal blood vessels (Bengiamin et al. 2009).

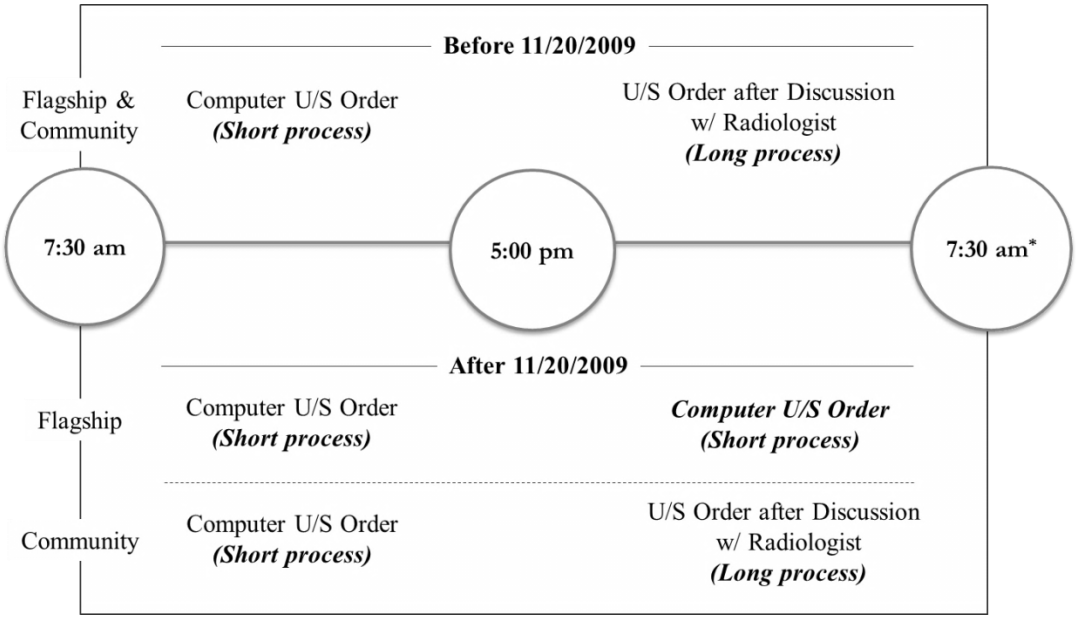
U/S is one of many diagnostic modalities available to the ED physician for patients with abdominal pain, thus supporting that this is a high discretion setting and the availability of a specific test might affect the demand for the test. Specifically, in addition to U/S, other common diagnostic modalities available to the ED physician include laboratory testing, and other types of imaging such as Computed Tomography (CT), or X-rays. Unlike U/S, these other tests are consistently available to the ED physician in our study group at all hours and on weekends and can be directly ordered via the electronic medical record (EMR) system.

Our interest is in how removing barriers for U/S ordering affects use. We exploit a process change that occurred at Flagship and Community: the process used by ED physicians to order an U/S was changed at Flagship and not Community. For an U/S to be completed, a radiology technician must first perform the scan, after which, the radiologist reviews the scan and interprets it. Before the process change, ED physicians who ordered an U/S at night or on the weekends at both hospitals had to discuss the particular case with the radiologist and request authorization for the study. If the radiologist approved it, then the U/S technician was called in to the hospital to perform it. In contrast, during the day, there was an U/S technician on duty in the hospital and U/S were ordered directly by the ED physician through the computerized ordering system, and no prior authorization was required.

In August 2009—which is before the September start date of our dataset—Flagship (but not Community) changed to 24/7 technician coverage so that an in-house technician would be available at night. The change was in part a response to staffing challenges that arose from having the technicians cover the overnight shift on an on-call basis: hospital and regulatory requirements restricted technicians from working the next day if they were called in at night. On November 20, 2009, Flagship switched their U/S order process so that ED doctors could place a direct order for an U/S on weekends and in the evening hours from 5:00 pm-7:30 am using the same computerized ordering system that they used during the day (See **Figure 3.3.1**). Prior to this change, physicians at Flagship had to complete a manual paper

order for the U/S study after it was authorized by the radiologist. The net effect was that the total time required for a physician to order an U/S at Flagship decreased from 10-15 minutes to less than one minute. At Community, however, there was no change in the U/S order process. For our purposes, the reduced time to order an U/S on nights and weekends at Flagship is equivalent to increasing the physicians' capacity to order U/S because it decreased the time it took them to place an order.

Figure 3.3.1. Timeline of Changes in U/S Ordering Process at Flagship and Community



Change in Ordering Process November 20, 2009
**Weekends follow same procedure as night shift*

3.4 HYPOTHESES

The primary driver of whether or not the physician orders an U/S is the physician's clinical concern for a particular subset of emergent diagnoses. However, the clinical guidelines about U/S and diagnostic test usage are not clear cut, thus resulting in significant inter-physician variability in the number of tests that they order (Stiell et al. 1997). In addition, operations management research suggests that additional variability may be introduced by non-medical factors. For example, previous research has shown that physicians' medical decisions can be influenced by their environment, such as how many patients are in the hospital unit (Batt et al. 2012, KC and Terwiesch 2012, Kuntz et al. 2013).

In this paper, we consider another state-specific factor that may influence the ordering of U/S: how long it takes for a physician to place the order. There is a speed-quality tradeoff when performing tasks in service settings: spending more time performing a task for a customer increases the quality of work for that customer, but delays other customers, thus reducing their quality of service (Anand et al. 2011,

George and Harrison 2001, Stidham and Weber 1989). This speed-quality tradeoff changes based on the state of the system, and workers adapt their behavior accordingly. Specifically, Hopp, Iravani and Yuen (2007) theorize that in DTC settings, quality can serve as a buffer for workload variability. When employees have a light workload, they are able to spend more time with their customers, which results in a higher quality of service. Conversely, if workers have a high workload, they may be forced to spend less time with each customer, which decreases quality, but enables more customers to be served. Moreover, if the time to complete a task is reduced, then the optimal speed-quality combination shifts since it requires less time to provide the same level of quality, and consequently, more services will be provided (Hopp et al. 2007). In the hospital, because ED physicians provide care to multiple patients simultaneously, they face the tradeoff that spending more time on one patient's care means less time is available for other patients. Thus, if a process improvement makes a task quicker to perform the physician can achieve higher quality performance for one patient for a smaller cost to his or her other patients.

Flagship's U/S process change removed a step in the process, which reduced the time it took physicians to order an U/S, increasing the theoretical maximum number of U/S that ED physicians could order during their shift (Cachon and Terwiesch 2004). Hopp, Iravani and Yuen (2007) state that increasing capacity will encourage workers to increase the quality of service by providing additional services to customers. A different way of framing this is that the process change reduced physicians' marginal cost of ordering an U/S, thereby increasing the likelihood that physicians would choose to order U/S for more of their patients. Performing additional tests has been shown to improve a physician's confidence in her diagnosis, a desirable outcome for physicians (Abramson et al. 2000). An additional motivation for physicians to order imaging more generally is that patients are more satisfied with their experience when more diagnostic tests are performed (Sun et al. 2000). Therefore, given the reduced marginal cost of ordering an U/S, and the benefits to the physician and patient for having one performed, we expect that after the change in the U/S ordering process, there is a higher probability of an U/S being ordered, all else being equal.

HYPOTHESIS 1 (H1): *The probability a patient will receive an U/S after the change in U/S ordering procedures will be greater than for an equivalent patient before the change in the ordering process.*

Although the change in the U/S ordering process reduces the amount of time to order each U/S, patients could end up staying in the ED longer after the process change than before if the reduced marginal cost causes the physician to order an U/S that he or she otherwise would not have ordered under the old process. In general, an U/S study takes between 15-30 minutes to perform depending on the specific study ordered. In addition, there may be queuing and transport time before the study, adding to

the study time. Once completed, the radiologist has to review the study and provide an interpretation for the ED physician, which takes additional time. Therefore, patients who receive an U/S have a longer ED stay than patients who do not have an U/S, all else being equal. The longer stay due to the additional U/S contributes to congestion in the ED, consistent with Hopp, Iravani and Yuen's (2007) prediction that adding capacity will result in increased congestion because servers will respond to the additional capacity by providing more services to their customers, thus increasing service time. Moreover, the increase in U/S places a higher workload on radiologists. Radiologists read U/S, x-rays, and CT scans, thus their services are shared across a wide range of ED patients. Our research setting is therefore slightly different from the setting in Hopp, Iravani and Yuen (2007), which assumes only one customer type in their model, and does not address the impact of increased service provisions on shared resources. Other research has shown that when services are shared, additional demand of one type can create congestion for all types of customers, or in this case, patients sharing that service (Berry Jaeker and Tucker 2014, Chao et al. 2001). Consequently, in our study, we expect that the increase in the number of U/S will delay radiological test results for all patients. We predict that the average LOS for ED patients will be longer after the change due to the cumulative effect of the additional time required to perform the increased number of U/S, and the increased probability of queuing for radiology services.

HYPOTHESIS 2 (H2): *The average LOS of an ED patient after the change in U/S ordering procedures will be greater than for an equivalent patient before the change in the ordering process.*

The underlying objective of process improvement changes, such as the elimination of a step in the U/S ordering process in our study, is to improve service by reducing delays in care. This could also reduce the need to expand the ED to accommodate increasing demand. In the ED, if all the beds are full, arriving patients must wait in the waiting room, delaying the start of their care, which may significantly worsen outcomes (Pines and Hollander 2008). Therefore, an important measure of a process improvement is a change in the waiting time for service. If the ED is capacity constrained, a longer average LOS for patients will increase the waiting time for patients entering the ED. The increased congestion due to waiting patients is equivalent to the increased congestion described in Hopp, Iravani and Yuen's (2007) model. The congestion occurs despite reduced service time for the service of interest because customers' *total* service times become longer. Most EDs are facing increased demand for their services (Kellermann 2006), making it likely that the EDs in our study will be capacity constrained and therefore will experience an increase in waiting time. Thus, in our study, we would expect that because the LOS of ED patients increases after the change in U/S ordering procedure, this will result in a higher probability of waiting, and an increase average waiting time for patients to the ED.

HYPOTHESIS 3 (H3): *The waiting time of an ED patient after the change in U/S ordering procedures will be greater than for an equivalent patient before the change in the ordering process.*

One of the conclusions of Hopp, Iravani and Yuen (2007) is that while the reduced processing time for a task results in more workers deciding to provide that extra service to their customers, which then leads to congestion, this process change may nevertheless be optimal because the financial benefits that the company reaps from the increased quality outweigh the costs of the additional waiting time. Our study differs from Hopp, Iravani and Yuen because the EDs treat multiple patient types and therefore, to evaluate the net effect of reducing the U/S order processing time, we must account for its impact on all ED patients, not just at the individual patient type level. For example, any benefit for the patients who receive an U/S must be compared to the costs incurred by those patients who do not. Given this, we calculate the magnitude of the benefits for the patients who were at risk of an U/S after the processing time changes. Two ED clinical quality measures are admission rate to the hospital and the percentage of ED patients who are readmitted to the ED within 72 hours.

First, we predict that increasing U/S capacity will decrease admissions to the hospital. Additional U/S capacity will enable physicians to order U/S for patients they previously would not have, providing more information for the physician to be confident in her decision that the patient does not have a serious condition and can be safely discharged to home from the ED. Without easy access to an U/S, the physician might err on the side of admitting the patient to the hospital for further observation and testing.

HYPOTHESIS 4A (H4A): *The probability of admission for an ED patient who after the change in U/S ordering procedures will be smaller than for an equivalent patient before the change in the ordering process.*

Using the same underlying logic, we expect a decrease in the readmission rate. Because we anticipate that having an U/S results in a more accurate diagnosis, the patients should receive the necessary treatment to address the medical condition. Similarly, we also expect a smaller chance of a serious condition being "missed." These two results should reduce the probability that patients who have U/S will return to the ED within three days.

HYPOTHESIS 4B (H4B): *The probability of readmission to the ED within 72 hours for an ED patient after the change in U/S ordering procedures will be less than for an equivalent patient before the change in the ordering process.*

3.5 DATA AND ECONOMETRIC SPECIFICATION

3.5.1 DATA

Our data consists of patient level records for all adult ED patient visits at Flagship Hospital and Community Hospital, with a chief complaint of abdominal pain between September 30, 2009 and May 31, 2012. The beginning date of our data set is after Flagship began staffing a radiology technician in the hospital 24 hours a day, seven days a week, but before the change in the ordering process. Since our sample is restricted to abdominal pain patients, each patient has the potential to receive an U/S. For each visit, we have the patient's medical record number (MRN), demographic information (e.g. age, gender, race, insurance), emergency severity index (ESI, a measure of the patient's severity and urgency), treating physician, time of arrival, and length of service (see **Figure 3.5.1** for descriptive statistics). In addition, each record has time stamped medical and procedure orders, such as medication, laboratory, and radiology orders, as well as the reason for a patient's visit (known as the primary complaint), the final diagnosis, and the disposition (e.g. admission, discharge, transfer). We also have the total number of patients seen in each ED each day. In total, we have 17,773 unique patients with 25,149 patient visits in our study period.

Figure 3.5.1: Summary statistics

Hospital	**Flagship**		**Community**	
Abdominal Patient Visits				
*Before Change**	429		355	
After Change	9404		6950	
Beds	99**		42	
Average daily U/S performed:				
Weekday				
Before Change	1.098	(1.043)	0.761	(0.929)
After Change	1.485	(1.270)	0.834	(0.886)
Night/Weekend				
Before Change	0.429	(0.587)	0.344	(0.538)
After Change	1.603	(1.385)	0.468	(0.702)
Primary Insurance (% of Total)				
Private	50.5%		62.3%	
Medicare	11.0%		11.5%	
Medicaid	21.0%		15.2%	
Free care/Uninsured	14.8%		10.0%	
Other	2.7%		1.0%	
Race (% of total)				
White	71.8%		60.2%	
Black	15.0%		17.5%	
Other	13.2%		22.3%	
Severity Score (% of total)				
ESI 2	21.0%		4.4%	
ESI 3	72.0%		92.4%	
Other	7.0%		3.2%	
Average LOS in ED (Service time, in minutes)	267.9	(120.8)	253.2	(114.6)
Average Wait time to enter ED (in minutes)	90.2	(93.6)	30.2	(46.5)
Average ED Occupancy	83.1%	(7.64)	79.2%	(8.87)
% Admitted to hospital	23.2%		25.3%	

% Readmitted		
Within 72 hours	2.26%	2.53%
Within 7 days	4.12%	4.46%
% of cases Resident present	79.2%	12.3%

*Before Change = September 20, 2009-November 20, 2009; After Change = November 21, 2009- May 31, 2012; **Includes 20 psychiatric beds; Std. deviations in ( )

Since we are interested in a change in policy at one hospital that did not occur at the other, we use a difference-in-differences methodology to study the effects of the change (Angrist and Pischke 2009). The benefit of this type of model is that it controls for any unobservable trends at the hospitals. However, it also requires that the "parallel trends" assumption is met (Angrist and Pischke 2009). Specifically, other than the change of interest, both sites should have the same trend in the outcome variable over time and with no other changes during the time of the study in one, but not the other location. Therefore, we restrict our sample to only those patients who saw a physician who practices at both hospitals, thus eliminating differences between the two locations due to differences in physician behavior. We also confirmed that there was no change in patient characteristics between the two EDs over time. To do this, we ran a set of t-tests, Wilcoxon-Mann-Whitney, Chi-square, and Fisher's exact tests to compare the age, gender, ESI, and race of patients at each hospital across time (see **Figure A.2** in the **Appendix**). None of the tests were significant at the 0.05 level. Thus, we find no evidence that either hospital had a statistical change in patient types after the processing change, thus supporting that the parallel trends assumption is valid in our study.

3.5.2 MAIN MODEL FOR HYPOTHESIS 1

Hypothesis 1 predicts that Flagship night and weekend patients will have a higher probability of having an U/S after the process change than before. To test this hypothesis, we model the probability of U/S as follows:

$$Pr(\mathrm{U/S_{i,h}}) = \beta_0 + \beta_1 \mathrm{Controls_i} + \beta_2 \mathrm{PrimComplaint_i} + \beta_3 \mathrm{Flagship_i} + \beta_4 \mathrm{NightWeekend_i} + \beta_5 \mathrm{PostNW_i} + \beta_6 \mathrm{Flagship_i} * \mathrm{PostNW_i} + \varepsilon_{i,h} \quad (3.1)$$

$Pr(U/S_{i,h})$ is the probability of an U/S for patient i at hospital h. Since our variable of interest is binary, we use robust logistic regression, with standard errors clustered by physician (in all models), to predict the probability that a patient will receive an U/S. We employ a difference-in-differences methodology, taking advantage of the process change that occurred at Flagship. To do this, we control for the underlying differences between Flagship and Community with a binary variable $Flagship_i$, which is equal to one if the patient is treated at Flagship. In addition, we control for any trend that occurs at both hospitals after the

processing change. However, the U/S ordering change only directly affected some patients, specifically those who arrived on the weekend or at night, therefore we control for the fact that these patients are predicted to have a different effect from the weekday patients. To account for these patients, we introduce $PostNW_i$, which is a binary variable equal to one if patient i is treated at night or on the weekend after the change in U/S policy. We define night as arrival between 5pm and 5am because this corresponded to the times during which physician had to get radiologist approval before the change in the U/S ordering policy. It should be noted that the actual night hours went until 7:30 am, but we truncate the night definition as a patient who arrives closer to 7:30 am has a high probability of overlapping with the daytime hours. $Flagship*PostNW_i$ is a binary variable equal to one if patient i arrives to Flagship at night or on the weekend after the U/S policy change. We are interested in β_6, the coefficient on $Flagship*PostNW_i$, as this represents the additional probability of ordering an U/S for patients who were affected by the U/S policy change.

We control for several observable variables, while underlying patient physiological differences are captured in our error term. $Controls_i$ is a vector of control variables that includes age, age squared, primary and secondary insurance, race, gender, arrival at night or on the weekend, the quarter of the year (for seasonal effects), and year (for trend). Given the high autonomy of physicians, as well as their highly variable training styles and personal skills, there can be large variation in practicing styles, particularly in the use of tests and medications (Stiell et al. 1997). Therefore, in addition to the primary complaint and demographics of a patient, we control for the attending physician. We also control for the presence of a resident or mid-level provider as they are present on some, but not all, cases and may influence decision making.

We also control for the daily census of hospital h's ED in which patient i was treated. Previous work has shown that inpatient occupancy can affect the LOS of a patient and the use of resources (KC and Terwiesch 2012, Kuntz et al. 2013). To account for similar relationships between occupancy and resource use in the ED, we control for ED census. Since our EDs are different sizes, we convert the daily census into an occupancy percentage (between 0 and 100%) so that we can compare across our two study sites. ED capacity is not defined by the number of rooms because most EDs (including Flagship and Community), use hallway beds and boarding areas to provide flexibility in ED bed capacity. Therefore, following Kuntz, Mennicken and Scholtes (2013), we create an occupancy based on the census of each day divided by the maximum daily census, where the maximum daily census is the 99^{th} percentile of daily censuses.

Finally, $PrimComplaint_i$, is a vector of symptom variables, described in more detail in the section below, where each symptom is binary with one meaning patient i's primary complaint includes that

symptom and zero meaning it does not. This vector controls for the impact of the patient's underlying medical condition on the probability of receiving an U/S.

3.5.3 COMPLAINTS AND SYMPTOMS ASSOCIATED WITH ULTRASOUND

Some patients with abdominal pain are more likely to receive an U/S than others. Specifically, the likelihood of an U/S is dependent on the clinician's level of suspicion that the patient has a condition for which an U/S is considered diagnostic, which is dependent on the patient's symptoms, gender, age, and other characteristics. As an example, U/S is a first-line choice for disorders of the gallbladder, however it is not considered as useful for disorders of the colon. However, due to limitations of the EMR, we do not have access to the particular disorders or conditions that the physician was considering when ordering specific tests or studies. Instead, we only have the patient's stated complaints, which are typically symptom-based rather than diagnosis-based. For certain complaints, such as "Right-upper quadrant abdominal pain," we might expect a higher likelihood of an U/S since right-upper quadrant abdominal pain is a symptom that is often associated with a disorder of the gallbladder.

To account for primary complaint in predicting U/S usage, we constructed an index, $PrimComplaint_i$, of patients' symptoms/complaints. We first compiled a list of conditions where U/S is considered a useful diagnostic modality, based on the American College of Emergency Physician's (ACEP) policy statement (American College of Emergency Physicians 2008) (see **Figure A.3,** column 1 in Appendix). Next, we used the MedlinePlus database, an online medical encyclopedia produced by the U.S. National Library of Medicine and the National Institutes of Health (2013) to provide a list of symptoms or symptom categories associated with the U/S sensitive diagnoses identified in step one (see **Figure A.3,** column 2 in Appendix). This approach resulted in 59 symptom categories (e.g., nausea, vomiting, and flank pain). In addition to typical symptoms, such as abdominal pain or a cough, a patient's primary complaint could also include any recent surgeries or history of past medical illnesses that may be relevant to the use of U/S. For example, if a patient complains of flank pain and they had previous kidney stones, they would be more likely to receive an U/S than a patient complaining of general abdominal pain. As another example, the primary complaint could contain a note about a suspected disease, such as appendicitis, that either a referring doctor or the triage nurse believes to be the cause of the symptoms. These examples indicate that for some of the patients there was additional patient information, which although not a physical symptom, was related to the conditions in the ACEP guidelines for which an U/S would be appropriate. Therefore, we incorporate this information by adding 12 additional "symptom" categories to the list of conditions where U/S is a useful diagnostic modality that account for any history or previous medical judgment related to the conditions in the ACEP guidelines. Finally, there are some symptoms or

complaints, such as alcohol abuse, which are not described in the ACEP guidelines, but could change the likelihood of an U/S being performed. These conditions add an additional nine categories. All 80 symptom categories used are shown in **Figure A.4** in the Appendix .

Each of our symptom 80 categories is a binary variable equal to one if the patient's primary complaint (which can include multiple symptoms) includes that symptom and zero otherwise. For example, a patient's primary complaint could be three symptoms: nausea, vomiting, and abdominal pain, which would be coded as one for each of those symptoms, and zero for the other 77. One can consider these symptom categories as analogous to comorbidity categorizations, such as Elixhauser's comorbidity measures (Elixhauser et al. 1998), to predict LOS and the probability of mortality within the hospital. In our sample, there were 5,509 unique primary complaints, with 1.54 U/S-related symptoms per patient on average, for a total of 38,722 symptom complaints. The first author categorized the symptoms associated with each of the 5,509 unique primary complaints into the 80 symptom categories described above. In our study, we used these variables to predict, in addition to patient demographics, the propensity for an U/S to be used.

There were numerous free-texted complaints that did not exactly match one of the clinical symptoms, but were nevertheless similar to one. To ensure that our results were not impacted by our interpretation of the free text symptoms, we went through the 38,722 symptoms and separated them into two categories: those that clearly fell into a symptom category and those that required interpretation. Almost all (99.8%) of the symptom complaints fell cleanly into one of the 80 symptom categories. The remaining 0.2% of symptoms was less clear, so we used the context to make our final categorization. For example, a primary complaint of "Constipation – cramping" does not explicitly say abdominal cramping, but given the context, we coded that symptom as abdominal cramping. The first author and a research assistant both coded the 77 symptoms that were less clear and achieved a 0.66 ($p<0.01$) kappa value, which indicates a substantial inter rater reliability (Landis and Koch 1977).

3.5.4 MODELS FOR HYPOTHESES 2, 3, AND 4

To test Hypothesis 2, we analyze whether the change in the U/S ordering process increased the LOS for evening and weekend Flagship Hospital ED patients. We measure LOS as the time from the start of care to the time care is completed (in minutes). In our regression, we use the log LOS, since LOS is exponentially distributed. We perform a robust OLS regression using the following model to test whether the process change impacted LOS.

$$\begin{aligned}\text{Ln(LOS)} = {} & \beta_0 + \beta_1 Controls_i + \beta_2 PrimComplaint_i + \beta_3 Flagship_i \\ & + \beta_4 NightWeekend_i + \beta_5 PostNW_i + \beta_6 Flagship_i * PostNW_i \\ & + \beta_7 U/S_i + \beta_8 U/S * NW_i + \beta_9 U/S * PostNW_i + \beta_{10} U/S \\ & * Flagship * PostNW_i + \varepsilon_{i,h}\end{aligned} \quad (3.2)$$

If Hypothesis 2 is supported, $\beta_6$will be positive and significant. In addition, since an U/S is likely to increase the LOSs for those patients who have one, we include a dummy variable that is one if the patient had an U/S. We also include variables for whether the U/S was performed on a night or weekend, before or after the change in process, and whether it was performed at Flagship.

As a result of the hypothesized increase in U/S use and service time in the ED, we predicted that ED patients' waiting time would be greater at Flagship after the process change (Hypothesis 3). We defined waiting time as the time between arrival/check-in and time brought to a bed. In our dataset, there are a significant number of patients with zero wait. Therefore, we use a count model, and since the variance is much greater than the mean, we use a negative binomial regression model. When a patient is waiting, he cannot have an U/S that would affect his waiting time, so we do not need to include that in our model, which leaves us with

$$\begin{aligned}WaitTime_{i,h} = {} & \beta_0 + \beta_1 Controls_i + \beta_2 PrimComplaint_i + \beta_3 Flagship_i \\ & + \beta_4 NightWeekend_i + \beta_5 PostNW_i + \beta_6 Flagship_i * PostNW_i \\ & + \varepsilon_{i,h}\end{aligned} \quad (3.3)$$

If Hypothesis 3 is supported, β_6 will be positive and significant.

We predicted that Flagship ED patients who receive an U/S after the process change will be less likely to be admitted to the hospital (Hypothesis 4a) and less likely to be readmitted to the ED within three days (Hypothesis 4b). More specifically, we have:

$$\begin{aligned}\Pr(Event_{j,i,h}) = {} & \beta_0 + \beta_1 Controls_i \\ & + \beta_2 PrimComplaint_i + \beta_3 Flagship_i + \beta_4 NightWeekend_i \\ & + \beta_5 PostNW_i + \beta_6 Flagship_i * PostNW_i + \varepsilon_{i,h},\end{aligned} \quad (3.4)$$

where $Event_{j,i,h}$ is event j occurring to patient i in hospital h, where j is admission to the hospital or readmission to the ED within three days. For the models described in Eq. 5, we again use a logistic regression model. The outcome variable is one if a CT (or admission or readmission) was performed, and zero otherwise. If H4a is supported, then β_6 will be negative and significant, and it will also be negative and significant if H4b is supported.

3.5.5 ADDITIONAL ANALYSES

We conduct additional analyses to deepen our understanding of the impact of the ordering policy change, as well as to run robustness checks. Given that we predict more U/S, we are interested in seeing if the additional load for radiologists affects the time it takes to complete other radiological studies, which can contribute to the change in LOS. To do this, we measure the time between a radiology test being ordered, excluding U/S, and when the results of the test are returned. If the patient had more than one radiological study, we take the maximum time for returning the results. As with our LOS model, we take the log of this value given the exponential distribution of the time to test return. We regress the controls and difference-in-differences variables on this logged time using a robust OLS model, to get the following:

$$\begin{aligned} &Ln\left(RadTestReturnTime_{i,h}\right) \\ &\quad = \beta_0 + \beta_1 Controls + \beta_2 PrimComplaint_i + \beta_3 Flagship \\ &\quad + \beta_4 NightWeekend_i + \beta_5 PostNW_i + \beta_6 Flagship_i * PostNW_i \\ &\quad + \varepsilon_{i,h} \end{aligned} \tag{3.5}$$

If the policy change impacts the time required for radiology to return test results, β_6 will be positive and significant.

To test if reducing the barrier for an U/S changes the use of other resources, we measure the number of laboratory tests ordered and the likelihood of having a CT ordered after the processing change. To measure the number of laboratory tests ordered for a patient, we use a count model, and since the variance is roughly equal to the mean, we use a Poisson regression. We control for whether the patient had an U/S or not as in eq. (3.4).

$$\begin{aligned} &\text{NumLabTests} \\ &\quad = \beta_0 + \beta_1 Controls_i + \beta_2 PrimComplaint_i + \beta_3 Flagship \\ &\quad + \beta_4 NightWeekend_i + \beta_5 PostNW_i + \beta_6 Flagship_i * PostNW_i \\ &\quad + \beta_7 U/S_i + \beta_8 U/S * NW_i + \beta_9 U/S * PostNW_i + \beta_{10} U/S * Flaghsip \\ &\quad * PostNW_i + \varepsilon_{i,h} \end{aligned} \tag{3.6}$$

If the policy change reduces the number of lab tests, β_6 will be negative and significant. For the CT probability, we use a logit model as follows:

$$\begin{aligned} &\Pr\left(Event_{j,i,h}\right) \\ &\quad = \beta_0 + \beta_1 Controls_i + \beta_2 PrimComplaint_i + \beta_3 Flagship \\ &\quad + \beta_4 NightWeekend_i + \beta_5 PostNW_i + \beta_6 Flagship_i * PostNW_i \\ &\quad + \varepsilon_{i,h} \end{aligned} \tag{3.7}$$

If the policy reduces CT use, β_6 will be negative and significant.

3.6 RESULTS

Figure 3.6.1 shows the results from an OLS regression testing the impact of the process change on the LOS of patients who received an U/S (eq. 3.2). We ran this equation first to verify that the process change reduced the time required for patients to receive an U/S. As shown in Model 1, we find that the LOS for Flagship ED patients who received an U/S during the night or weekend after the ordering process change was shorter (β = -0.210, p<0.01) than before the change. Since this is an OLS with a log transform of the dependent variable, this is equivalent to a 21.0% decrease in LOS for a patient who receives an U/S on the night/weekend after the change at Flagship, when compared to the same patient receiving an U/S an night/weekend before the change. This change represents a reduction of more than one hour in the LOS in the ED. Therefore, these results suggest that—given that a patient received an U/S before the change—the ordering process change reduced the ED LOS.

Next, we look if this reduction in processing time is associated with an increase in U/S orders at Flagship (Hypothesis 1). In **Figure 3.6.1**, in Model 2, the base model (not including the difference-in-differences time effects) that predicts the probability of an U/S, we control for patient characteristics and show that there is a significant increase in the number of U/S performed at Flagship (β= 0.659, p<0.01), with a patient at Flagship having a 7.9 percentage points higher predicted probability of receiving an U/S than patients at Community. When we refine our analysis to include controls for the change in U/S ordering process (Model 3), the coefficient for Flagship patients on nights and weekends after the change is significant and positive (β=0.965, p<.01), providing support for Hypothesis 1. The average marginal effect (AME) indicates that Flagship ED patients on nights and weekends after the ordering process change have an 11.5 higher percentage point probability of having an U/S ordered than patients at Community on nights and weekends after the change. The U/S ordering process change results in an increase in the average predicted probability of a night/weekend patient at Flagship receiving an ultrasound from 9.4% to 20.3%. This result confirms our expectation that when U/S are quicker to order, physicians order more of them.

Figure 3.6.1: Impact of process change on probability of U/S

	(1)	(2)		(3)	
	OLS Log ED LOS	*Probability of U/S Logit Base Model*		*Probability of U/S Logit Diff-in-Diffs*	
		Coefficient	AME	Coefficient	AME
Flagship	-0.059**	0.659**	0.079**	0.180*	0.021*
	(0.019)	(0.072)	(0.009)	(0.081)	(0.010)
Night/Weekend	0.049	-0.284**	-0.034**	-0.779**	-0.093**
	(0.036)	(0.057)	(0.007)	(0.200)	(0.024)
Night/Weekend	-0.133**	-	-	-0.141	-0.017

After Change	(0.032)	-	-	(0.221)	(0.026)
Flagship Night/Weekend	0.110**	-	-	0.965**	0.115**
After Change	(0.022)	-	-	(0.107)	(0.013)
U/S ordered	0.303**	-	-	-	-
	(0.013)	-	-	-	-
U/S ordered	0.159**	-	-	-	-
Night/Weekend	(0.051)	-	-	-	-
U/S ordered Night/	0.026	-	-	-	-
Weekend After Change	(0.052)	-	-	-	-
U/S ordered Flagship Night/	-0.210**	-	-	-	-
Weekend After Change	(0.030)	-	-	-	-
Constant	5.082**	-0.881*	-	-0.528	-
Controls	Yes	Yes	Yes	Yes	Yes
Primary Complaint	Yes	Yes	Yes	Yes	Yes
Number of Obs.	17,118	17,000	17,000	17,000	17,000
R^2	0.16	-	-	-	-
Adjusted R^2	0.15	-	-	-	-
Degrees of Freedom	49	-	-	-	-
Pseudo R^2	-	0.11	-	0.12	-

OLS model of the LOS within the ED, controlling for whether a patient receives an U/S (1). Logistic regression models of the probability of U/S, where the base model (2) is w/o the processing change and the diff-in-diffs model (3) is with it. The logistic regressions include regression coefficients and average marginal effects (AME). *Note*: Controls include gender, age, age^2, year, quarter of the year, attending, presence of a resident primary insurance, secondary insurance, ESI severity score and race; Robust standard errors clustered by attending in ( )
$^+p<0.1$; $*p<0.05$; $**p<0.01$

Hypothesis 2 predicted that the U/S policy change would be associated with an increase in ED LOS (eq. 3.2). Model 1 of **Figure 3.2.1** presents the effect on LOS before and after the change in policy on patients who receive an U/S (as described above), as well as those who do not. Model 1 shows that there was an 11.0% ($p<0.01$) increase in LOS, or around 26 minutes, within the ED for all Flagship abdominal pain patients seen at night or on the weekend after the change in the U/S ordering process, supporting Hypothesis 2. To explain how the average ED LOS increases when the time it takes to receive an U/S at Flagship decreases, we see that in general, receiving an U/S increases ED LOS by 30.3% ($p<0.01$), and an additional 15.9% ($p<0.01$) on the nights and weekends. Since more U/S are ordered after the change, the net effect on LOS is an increase.

In addition to analyzing the time spent in the ED for treatment, we also modeled the effect of the policy change on the waiting time of patients in the ED (eq. 3.3). Waiting time better reflects the impact of the process change on the ED's flow rate independent of whether the patient waiting ends up with an U/S. We present these results in **Figure 3.6.2**, Model 1. Given that we had to use a negative binomial count model, we cannot

Figure 3.6.2: Impact of process change on waiting times

	(1)	(2)
	Negative Binomial	*OLS Log Rad.*
	ED Wait Time	*Test Return Time*

	Coefficient	AME	Coefficient
Flagship	0.753**	49.241**	0.041
	(0.060)	(4.398)	(0.029)
Night/Weekend	0.235**	15.349**	0.288**
	(0.085)	(5.564)	(0.086)
Night/Weekend	-0.301**	-19.663**	-0.407**
After Change	(0.089)	(5.789)	(0.089)
Flagship Night/Weekend	0.400**	26.172**	0.271**
After Change	(0.073)	(4.700)	(0.035)
Occupancy	0.040**	2.632**	0.003*
	(0.003)	(0.187)	(0.001)
Constant	-0.808*	-	4.687**
Ln Alpha Constant	0.406**	-	-
Controls	Yes	Yes	Yes
Primary Complaint	Yes	Yes	Yes
Number of Obs.	17,121	17,121	6,859
R^2	-	-	0.09
Adjusted R^2	-	-	0.06
Degrees of Freedom	-	-	49

Negative binomial regression of wait time to enter ED (3). OLS regression predicting time to return radiology tests (4); Robust standard errors clustered by attending in ( )
$^{+}p<0.1$; *p<0.05; **p<0.01

easily interpret the effect sizes associated with the coefficients. Therefore, we also present the marginal effects for patients seen at Flagship on the nights and weekends after the change, and find that for these patients, the policy change results in an average predicted increase in waiting time from 52 to 78 minutes (p<0.01), in support of Hypothesis 3. To provide more insight into why the change in U/S ordering process increases ED LOS and waiting time, we also analyzed if there was a change in the time to return radiological tests (eq. 3.5) given that radiology services are shared among all ED patients. Our results, presented in Model 2 of **Figure 3.6.2** show that there is a 27.1% (p<0.01), or approximately 30 minute increase in the time it takes to return radiology tests (other than U/S) after the change, which is consistent in explaining the increased LOS and waiting times in the Flagship ED.

Next, we analyzed if the U/S policy change was associated with a change in clinical quality measures, as measured by admission to the hospital and readmission to the ED within 72-hours (we also include within 7 days for robustness) (eq. 3.4), and we present the results in **Figure 3.6.3**, Models 1-3. We do not find any statistical support for an improvement in clinical quality measures, as measured by a change in admission rate or readmission to the ED, and thus we cannot reject the nulls for Hypotheses 4a and 4b.

Figure 3.6.3: Impact of process change on clinical quality measures

	(1) *Logit Admission*		(2) *Logit Readmit: 72 hours*		(3) *Logit Readmit: 7 days*	
	Coefficient	AME	Coefficient	AME	Coefficient	AME
Flagship	-0.387**	-0.056**	-0.208	-0.005	-0.273^{+}	-0.011^{+}
	(0.075)	(0.011)	(0.207)	(0.005)	(0.144)	(0.006)

Night/Weekend	0.158	0.023	-0.781+	-0.019+	-0.731+	-0.030+
	(0.163)	(0.023)	(0.431)	(0.011)	(0.387)	(0.016)
Night/Weekend After Change	-0.119	-0.017	0.858+	0.021+	0.599	0.025
	(0.180)	(0.026)	(0.473)	(0.012)	(0.400)	(0.016)
Flagship Night/ Weekend After Change	0.032	0.005	-0.215	-0.005	0.093	0.004
	(0.090)	(0.013)	(0.192)	(0.005)	(0.168)	(0.007)
Occupancy	-0.005+	-0.001+	-0.004	-0.000	-0.009	-0.000
	(0.003)	(0.000)	(0.008)	(0.000)	(0.006)	(0.000)
Constant	-3.302**	-	-3.976**	-	-2.947**	-
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Primary Complaint	Yes	Yes	Yes	Yes	Yes	Yes
Number of Obs.	17,073	17,073	15,627	15,627	16,557	16,557
Pseudo R^2	0.19	-	0.07	-	0.05	-

Logistic regression models of the probability of Admission to the hospital (1), Readmission to the ED within 72 hours (2), and Readmission to the ED within 7 days (3); Robust standard errors clustered by attending in ( )
+p<0.1; *p<0.05; **p<0.01

Finally, to understand any other consequences as a result of the change in ordering process, we tested if there was a change in the number of laboratory tests performed (eq. 3.6, **Figure 3.6.4**, Model 1) or the probability of receiving a CT (eq. 3.4, **Figure 3.6.4 5**, Model 2). We find that the predicted number of medical laboratory tests for night/weekend patients at Flagship drops from 7.86 to 7.60 (p<0.01). However, we find that the average marginal effect of the ordering process change is to increase the probability of a CT scan by 4.5% (p<0.01) with the predicted probability increasing from 50.1% to 54.6%. We discuss the implications of these results in the next section.

Figure 3.6.4: Impact of process change on use of other medical tests

	(1) *Poisson Lab Test Count*		(2) *Logit CT Scan*	
	Coefficient	AME	Coefficient	AME
Flagship	-0.032**	-0.247**	-0.287**	-0.063**
	(0.011)	(0.087)	(0.074)	(0.016)
Night/Weekend	0.006	0.049	0.450**	0.098**
	(0.017)	(0.128)	(0.120)	(0.026)
Night/Weekend After Change	0.013	0.102	-0.488**	-0.106**
	(0.017)	(0.131)	(0.126)	(0.027)
Flagship Night/Weekend After Change	-0.033**	-0.254**	0.209**	0.045**
	(0.010)	(0.081)	(0.062)	(0.013)
Occupancy	0.000	0.001	-0.001	-0.000
	(0.000)	(0.003)	(0.002)	(0.000)
U/S ordered	0.076**	0.589**	-	-
	(0.010)	(0.078)	-	-
U/S ordered Night/Weekend	0.116*	0.898*	-	-
	(0.056)	(0.433)	-	-
U/S ordered Night/Weekend After Change	-0.111*	-0.860*	-	-
	(0.055)	(0.422)	-	-
U/S ordered Flagship Night/ Weekend After Change	0.005	0.037	-	-
	(0.023)	(0.174)	-	-

Constant	1.951**	-	-1.294**	-
Controls	Yes	Yes	Yes	Yes
Primary Complaint	Yes	Yes	Yes	Yes
Number of Obs.	17,138	17,138	17,119	17,119
Pseudo R^2	0.03	-	0.10	-

Poisson count model of number of laboratory tests ordered (1). Logistic regression model of the probability of a CT scan (2). Robust standard errors clustered by attending in parentheses.
$^{+}p<0.1$; *p<0.05; **p<0.01

3.7 ROBUSTNESS

In addition to showing that the change in U/S orders affects radiological test return time, which provides support that the increase in LOS was due to the processing change, we also performed a further robustness check to provide additional evidence that the changes in U/S orders were a function of the change in ordering process, and not due to some other underlying trend. Specifically, we re-ran the original logistic regression, restricting our analysis to patients seen in 2010 or later. We also interacted the Flagship*NW effect with the years 2011 and 2012, thus Flagship*NW in 2010 is the baseline effect. If the processing change explains the increase in U/S use, then the greatest difference in U/S use should occur in 2010 and thus show up in the baseline, with no other effect in 2011 and 2012. Our results confirm this; the only statistically significant change in U/S use occurred in 2010 (See **Figure 3.7.1**).

Figure 3.7.1: Impact of process change on U/S use across years

	Probability of U/S Logit Model	
	Coefficient	Std. Err.
Flagship	0.192*	(0.080)
Night/Weekend	-0.954**	(0.113)
Night/Weekend_2011	0.064	(0.119)
Night/Weekend_2012	0.081	(0.190)
Flagship Night/Weekend	0.851**	(0.116)
Flagship Night/Weekend_2011	0.158	(0.129)
Flagship Night/Weekend_2012	0.169	(0.203)
Constant	-0.100	(0.372)
Controls	Yes	-
Primary Complaint	Yes	-
Number of Obs.	15,562	-
Pseudo R^2	0.12	-

Logistic regression model of the probability of U/S, for patients seen during or after 2010.
Robust standard errors clustered by attending in ( ).
$^{+}p<0.1$; *p<0.05; **p<0.01

Finally, to provide further evidence that making U/S easier to order increased the probability that physicians would order low efficacy U/S, we tested whether there was a difference in propensity to order an U/S between patients whose symptoms were clearly linked to a need—or not—for an U/S versus for patients whose symptoms were more ambiguous as to the need of an U/S. Specifically, some complaints, such as abdominal cramping—the primary complaint for patients with cholecystitis—and pelvic pain—for ectopic pregnancies—are conditions which strongly indicate the need for an U/S. Other conditions, such as fainting, are clear in not needing an U/S. The left hand column of **Figure 3.7.2** contains eight such symptoms that are clear about the need, or not, for an U/S. Conversely, other complaints, such as abdominal pain, are more ambiguous about whether or not an U/S is warranted. For example, abdominal pain is a symptom for many conditions, most of which do not require an U/S, but can also be associated with cholecystitis or other acute conditions which warrant an U/S. The right hand side of **Figure 3.7.2** lists eight symptoms for which the need for an U/S is ambiguous.

To test for a difference in the use of U/S between patients with clear versus ambiguous symptoms for an U/S, we run logistic regressions for each primary complaint of interest on only those patients seen after the change, and we look at the effect of being seen at Flagship compared to Community. **Figure 3.7.2** shows that the coefficients for Flagship are not significant for the patients whose symptoms clearly warrant an U/S (or not, as in the case of the shaded symptoms). In contrast, the coefficients for Flagship are significant for the ambiguous symptoms. These findings support our explanation that ambiguous cases—which are "low efficacy uses of U/S"—are responsible for the increase in use of U/S at Flagship after the order process change.

Figure 3.7.2: Impact of process change on U/S use across complaints

Stable, Frequent (Rare) Use of U/S				**Varying Use of U/S**			
Complaint	*Propensity for U/S at Flagship*		*N=*	*Complaint*	*Propensity for U/S at Flagship*		*N=*
	Coefficient	Std. Error			Coefficient	Std. Error	
Abd. Cramps	0.790	(0.565)	103	Abd. Pain	0.768**	(0.053)	12,300
Vaginal Bleeding	-0.182	(0.632)	53	Flank (side) Pain	0.784**	(0.190)	806
Pelvic Pain	0.506	(0.437)	93	Nausea	0.773**	(0.127)	2,490
Ascites	2.025+	(1.073)	157	Vomiting	0.788**	(0.128)	2,602
Swelling (not abd.)	1.625	(1.087)	71	Back Pain	0.844**	(0.221)	692
Biliary Indication	0.154	(0.759)	66	Chest Pain	0.580**	(0.221)	806
Fainting/ Syncope	-0.470	(0.642)	102	Rt Upper Quad Abd Pain	1.073*	(0.441)	95
Abnormal Stool	0.363	(0.712)	149	Fever	0.604*	(0.295)	346

Logistic regression of U/S propensity by primary complaint comparing Flagship and Community after processing change. Note: All coefficients are for a logistic regression predicting U/S use; Conditions in gray rarely associated with U/S use; +p<0.1, *p<0.05, **p<0.01

3.8 DISCUSSION

We examined the impact of a change in the ordering process for U/S on nights and weekends for ED patients with abdominal pain at two hospitals within the same health system and staffed by the same ED physicians between 2009 and 2012. We found that the change—which decreased the time it took for ED physicians to order an U/S—decreased the LOS for patients who receive an U/S by 21%, or about an hour. However, the reduction in U/S processing time was associated with an 11.5 percentage point increase in the probability of a patient receiving an U/S. Although the processing change reduced the LOS of patients who had U/S, having an U/S compared to not having an U/S increased the LOS of ED patients by 30% with an additional 15.9% increase when ordered on the night or weekends. Thus, the cumulative effect was a net increase in the LOS in the ED. Furthermore, the expected waiting time for patients entering the ED increased by approximately 26 minutes. Although part of the increased LOS was from the additional U/S that were ordered, we also found that after the process change there was a delay in getting other radiological test results, such as CT, back from the now busier radiologists. Specifically, we found that the time to return non-U/S radiological tests increased by 27.1%, or approximately 30 minutes.

Unfortunately, it does not appear that patients' clinical quality measures improved as a result of the additional U/S. More specifically, we did not see a change in the admission rates of ED patients to the hospital or in the 3 or 7-day readmission rates to the ED. A curious finding was that the number of CT scans in Flagship on nights and weekends after the process change increased by 4.5%. Overall, the number of CT scans over time was decreasing, with a 10.6 percentage point decrease in the probability of receiving one on the night or weekend after the change, suggesting that in general, physicians were not ordering more scans. Instead, we suspect that physicians used CT scans in combination with U/S to be sure of an appropriate diagnosis if the U/S results were uncertain or negative, because if the leading explanation for the patients' symptoms were ruled out by the U/S, physicians might need a CT to understand what else might be causing the patients' symptoms.

Combined, these results have major implications for the research and implementation of operational improvements in DTC settings. Prior research and policies have focused on removing waste, either physical or labor, from a system in order to improve its performance. However, we find that removing what appears to be a wasteful step in a process (i.e., getting approval for a test from an additional doctor) actually creates additional inefficiencies in the system. These results suggest that behavioral responses to a system must be incorporated when trying to improve the efficiency of a system.

3.8.1 IMPLICATIONS FOR RESEARCH

Our study empirically validates that increasing servers' capacity to provide a service could result in an overall increase in congestion because workers with discretion over their tasks will use their additional capacity to provide more services to their individual customers. The additional services that are provided to customers increases their service time, which makes incoming customers' waits longer. Moreover, we show that the increase in service time also spilled over to patients who did not receive an U/S due to a shared external resource, radiology. This shared resource creates interdependencies between all patients who have any kind of radiological test. Our work shows that the interdependency between servers' decisions and the load on the shared resource causes an increase in service time to patients who do not actually receive the additional service. Given the occurrence of this amplification of cost across patients, the effects must be included when considering the cost versus benefit of process changes that increase capacity.

Another contribution is that our study includes the impact of incentives on discretionary behavior. In our paper, the incentives of the physicians are not necessarily in alignment with the incentives of the hospital. Specifically, a physician orders a test, such as an U/S, to improve the certainty of her diagnosis for her patient. However, the hospital is incentivized to increase throughput, while ensuring an appropriate level of care (i.e. avoiding a costly readmission to the ED, which is a sign that a diagnosis was "missed'), and thus might not want the physician to order additional U/S if they increase waiting times and LOS without increasing quality metrics.

Finally, our paper contributes to the literature on cost efficiency in healthcare by including a non-financially-driven motivation that explains why physicians order medical interventions that do not improve their patients' health. By decreasing the time required for a physician to order an U/S, the marginal cost to the physician of ordering an U/S is reduced, while the benefits to the physician remain the same. Specifically, the additional information provided by the U/S provides information to the physician which helps her diagnosis her current patient, and may bring additional satisfaction to a patient who generally believes more medical care is better. However, it is likely difficult for physicians to perceive that their higher rate of ordering U/S places an additional load on radiology that ultimately decreases all patients' experiences in the ED by increasing their wait times and LOS without providing a noticeable benefit in clinical quality measures.

3.8.2 IMPLICATIONS FOR PRACTICE

Our results have significant implications for practice. As we described above, we find that at least in an ED setting, a process improvement that reduces the processing time for providing a particular service can actually increase demand for that service, which results in increased congestion in the system and longer overall throughput time. Our study highlights an important lesson for process improvement: increasing process capacity at one step in a service delivery process can change the demand for that service in discretionary settings, and can even decrease performance by further overloading a downstream bottleneck resource that has to process the larger volume of demand (in our setting, radiology). Therefore, improvement initiatives should seek to optimize performance at the system—rather than local—level. Although it was not implemented as a process improvement project, a similar dynamic occurred when Starbucks introduced the time-consuming, but popular Frappuccino beverage without adding worker capacity. As a result, waiting times for all customers sharply increased, driving away customers who ordered drinks with shorter processing times (e.g., espresso), which reduced overall revenue (Adamy 2006). Another example is related to transportation. In many metro areas, highway congestion contributes to pollution and wasted worker productivity. Although a reasonable solution would seem to be to widen roads to increase highway capacity, state transportation department officials recognize that increasing capacity on roadways will encourage more people to drive on these roads, quickly causing congestion again (Emmett Brady 1993). Therefore, the net result will be the same high congestion, but this time at a higher load which causes even higher levels of pollution and more people stuck in gridlock. This example helps illustrate an important concept for practitioners: it may be optimal to have longer service times if this reduces overall demand, preventing additional costly delays in service. In our ED setting, it may be optimal to have a less efficient U/S ordering process with radiology as a gatekeeper to minimize low efficacy U/S orders. Given that clinical quality measures did not increase with the additional U/S orders, our results suggest that while the original process was more difficult for the ED attendings, when necessary, patients still received an U/S.

3.8.3 Limitations

As with any study, ours has limitations, which we have done our best to address. First, our dataset is limited in the pre-process change period due to a change in the hospital's data collection software that reduced the availability of data. We are therefore unable to separately analyze the effect of the increased availability of the technician. Second, we only have data from abdominal pain ED patients and therefore cannot comment on the impact of the process change on all ED patients. However, there is no reason to think that abdominal pain patients would have a different ED wait time than non-abdominal pain patients. Third, we control for physician effects to account for inter-physician variability. However, it is possible that a patient's care spreads across two physician shifts. In these situations, we used the first

attending assigned to the patient, who is typically the one responsible for the care plan, including the orders and disposition. If a second attending physician was involved, the standard practice for this physician group is that the second physician would do their best to execute the care plan as originally conceived. There may be instances where the physician to whom the patient was "signed-out" to may exercise discretion in changing the plan. We are unable to identify the frequency of this but we believe it is not significant based on observing the practice habits of the physician group. Fourth, we do not study the financial cost to insurance company and profit to hospital associated with the change in resource use. We recognize that Flagship might benefit from the additional payment from U/S, and these profits might outweigh the cost on LOS. However, the ordering physicians do not benefit financially from increased U/S as they are independent of the hospital and the radiology group who are paid for the service. Additionally, the ED physicians at both Flagship and Community are salaried with no change in compensation for additional tests ordered. In addition to financial benefit to the hospital, we are unable to assess if the additional U/S increase patient satisfaction, which might bring more future revenue to the hospital. We leave it to future research to examine these effects more closely. Finally, we recognize that our study is limited to two EDs at one health system. We have tried to control for any population factors, such as insurance, age, and other demographics, but of course we cannot prove that these results replicate at other institutions. Nevertheless, we feel that results have significant implications for both research and practice.

3.9 CONCLUSIONS

Our work empirically shows that increasing resource capacity in an ED *increases*—rather than decreases—throughput time due to an increase in resource use to provide additional service. These results highlight the importance of accounting for endogenous changes in demand due to capacity changes in service settings, and suggest that due to behavioral responses to resource availability, what appears to be a wasteful step may not actually be inefficient. In healthcare, this is very important as our results provide an explanation for some of the ever increasing costs. Furthermore, we show that in the complex, interconnected system or hospitals, changes in resource capacity impact not just the patients who receive the additional resources, but other patients who share a resource. Our study suggests an operations-based solution of increasing the cost/difficulty of ordering discretionary but sometimes low-efficacy treatments to address the rise in healthcare spending. We show that, paradoxically, to improve hospital performance, it could be optimal to put into place "inefficiencies" to curb the desire to increase service that does not actually improve outcomes.

References

Abramson, S., N. Walders, K. E. Applegate, R. C. Gilkeson, M. R. Robbin. 2000. Impact in the emergency department of unenhanced ct on diagnostic confidence and therapeutic efficacy in patients with suspected renal colic. *American Journal of Roentgenology*. **175**(6) 1689-1695.

Adamy, J. 2006. *Starbucks earning rise 16%; wait time curbs sales growth*. New York City, NY.

Aiken, L. H., S. P. Clarke, D. M. Sloane, J. Sochalski, J. H. Silber. 2002. Hospital nurse staffing and patient mortality, nurse burnout, and job dissatisfaction. *J. Amer. Medical Assoc.* **288**(16) 1987-1993.

American College of Emergency Physicians. 2008. *Emergency ultrasound guidelines*. Dallas, TX.

Anand, K. S., M. F. Paç, S. Veeraraghavan. 2011. Quality–speed conundrum: Trade-offs in customer-intensive services. *Management Sci.* **57**(1) 40-56.

Anderson, D., B. Golden, W. Jank, E. Wasil. 2012. The impact of hospital utilization on patient readmission rate. *Health Care Manag Sci*. **15**(1) 29-36.

Anderson, D., C. Price, B. Golden, W. Jank, E. Wasil. 2011. Examining the discharge practices of surgeons at a large medical center. *Health Care Manag Sci*. **14**(4) 338-347.

Angrist, J. D., J.-S. Pischke. 2009. *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press, Princeton.

Argo, J. L., C. C. Vick, L. A. Graham, K. M. F. Itani, M. J. Bishop, M. T. Hawn. 2009. Elective surgical case cancellation in the veterans health administration system: Identifying areas for improvement. *The American Journal of Surgery*. **198**(5) 600-606.

Armony, M., I. Gurvich. 2010. When promotions meet operations: Cross-selling and its effect on call center performance. *M&SOM*. **12**(3) 470-488.

Batt, R. J., C. Terwiesch, O. A. Soremekun. 2012. Docs under load: An empirical study of state-dependent service rate mechanisms. *Working Paper, Wharton School of Business, Pennsylvania*.

Bengiamin, R., G. R. Budhram, K. E. King, J. M. Wightman. 2009. *Abdominal pain*. Mosby Elsevier, Philadelphia, PA.

Berry Jaeker, J., A. L. Tucker. 2014. Hurry up and wait: An empirical study of the spillover effects of workload on patient length of stay. *Harvard Business School Working Paper*.

Berwick, D. M., A. D. Hackbarth. 2012. Eliminating waste in us health care. *JAMA*. **307**(14) 1513-1516.

Best, T., D. D. Eisenstein, D. O. Meltzer, B. Sandıkçı, . 2013. Efficient management of strained inpatient bed capacity. *Working Paper*.

Blow, O., L. Magliore, J. A. Claridge, K. Butler, J. S. Young. 1999. The golden hour and the silver day: Detection and correction of occult hypoperfusion within 24 hours improves outcome from major trauma. *J. Trauma-Injury, Infection, and Critical Care*. **47**(5) 964.

Boudreau, J., W. Hopp, J. O. McClain, L. J. Thomas. 2003. On the interface between operations and human resources management. *M&SOM*. **5**(3) 179-202.

Buist, M. D., G. E. Moore, S. A. Bernard, B. P. Waxman, J. N. Anderson, T. V. Nguyen. 2002. Effects of a medical emergency team on reduction of incidence of and mortality from unexpected cardiac arrests in hospital: Preliminary study. *BMJ*. **324**(7334) 387-390.

Cachon, G. P., C. Terwiesch. 2004. *Matching supply with demand: An introduction to operations management*. McGraw Hill, Boston, MA.

California Office of Statewide Health Planning and Development. 2010. *Readmissions to california hospitals, 2005-2006*.

Centers for Medicare & Medicaid Services Office of the Actuary National Health Statistics Group. 2012. *National health care expenditures data*. https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/downloads/tables.pdf.

Chalfin, D. B., S. Trzeciak, A. Likourezos, B. M. Baumann, R. P. Dellinger. 2007. Impact of delayed transfer of critically ill patients from the emergency department to the intensive care unit *Crit. Care Medicine*. **35**(6) 1477-1483.

Chan, C. W., V. F. Farias, N. Bambos, G. J. Escobar. 2012. Optimizing intensive care unit discharge decisions with patient readmissions. *Oper. Res.* **60**(6) 1323-1341.

Chao, C., J. Zhanfeng, P. Varaiya. 2001. Causes and cures of highway congestion. *Control Systems, IEEE*. **21**(6) 26-32.

Clarke, A. 1996. Why are we trying to reduce length of stay? Evaluation of the costs and benefits of reducing time in hospital must start from the objectives that govern change. *Quality in Health Care*. **5**(3) 172.

Dallery, Y., S. B. Gershwin. 1992. Manufacturing flow line systems: A review of models and analytical results. *Queueing Systems*. **12**(1) 3-94.

Debo, L. G., L. B. Toktay, L. N. van Wassenhove. 2008. Queuing for expert services. *Management Sci.* **54**(8) 1497-1512.

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

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

Emmett Brady, M. 1993. Dynamic stability, traffic equilibrium and the law of peak-hour expressway congestion. *Transportation Research Part B: Methodological*. **27**(3) 229-236.

Falvo, T., L. Grove, R. Stachura, D. Vega, R. Stike, M. Schlenker, W. Zirkin. 2007. The opportunity loss of boarding admitted patients in the emergency department. *Acad. Emerg. Medicine*. **14**(4) 332-337.

Ferrand, Y., M. Magazine, U. Rao. 2010. *Comparing two operating-room-allocation policies for elective and emergency surgeries*.

Fisher, M., A. Raman. 1996. Reducing the cost of demand uncertainty through accurate response to early sales. *Oper. Res.* **44**(1) 87-99.

Freeman, M., N. Savva, S. Scholtes. 2014. Decomposing the effect of workload on patient outcomes: An empirical analysis of a maternity unit. *Working Paper*.

Fuchs, V. R. 2012. Major trends in the u.S. Health economy since 1950. *New England J. Medicine*. **366**(11) 973-977.

Gawande, A. 2007. *Better: A surgeon's notes on performance*. Metropolitan Books, New York, NY.

Gawande, A. 2009. The cost conundrum: What a texas town can teach us about health care. *The New Yorker* 36-44.

George, J. M., J. M. Harrison. 2001. Dynamic control of a queue with adjustable service rate. *Oper. Res.* **49**(5) 720-731.

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

Gino, F., G. P. Pisano. 2007. Toward a theory of behavioral operations (working paper).

Gittell, J. H., K. M. Fairfield, B. Bierbaum, W. Head, R. Jackson, M. Kelly, R. Laskin, S. Lipson, J. Siliski, T. Thornhill, J. Zuckerman. 2000. Impact of relational coordination on quality of care, postoperative pain and functioning, and length of stay: A nine-hospital study of surgical patients. *Medical Care*. **38**(8) 807-819.

Glouberman, S., H. Mintzberg. 2001. Managing the care of health and the cure of disease—part i: Differentiation. *Health Care Management Review*. **26**(1) 56-69.

Green, L., S. Savin, N. Savva. 2011. "Nursevendor problem": Personnel staffing in the presence of endogenous absenteeism. *Management Sci.*(Forthcoming).

Green, L. V. 2012. Om forum—the vital role of operations analysis in improving healthcare delivery. *M&SOM*. **14**(4) 488-494.

Green, L. V., V. Nguyen. 2001. Strategies for cutting hospital beds: The impact on patient service. *Health Services Res.* **36**(2) 421-442.

Green, L. V., S. Savin, B. Wang. 2006. Managing patient service in a diagnostic medical facility. *Oper. Res.* **54**(1) 11-25.

Guerriero, F., R. Guido. 2011. Operational research in the management of the operating theatre: A survey. *Health Care Manag Sci*. **14**(1) 89-114.

Gupta, D. 2007. Surgical suites' operations management. *Prod. & Oper. Management*. **16**(6) 689-700.

Ha, A. Y. 1998. Incentive-compatible pricing for a service facility with joint production and congestion externalities. *Management Sci.* **44**(12) 1623-1636.

Hand, R., P. Levin, A. Stanziola. 1990. The causes of cancelled elective surgery. *American Journal of Medical Quality*. **5**(1) 2-6.

Hasija, S., E. Pinker, R. A. Shumsky. 2010. Om practice—work expands to fill the time available: Capacity estimation and staffing under parkinson's law. *M&SOM*. **12**(1) 1-18.

Hauck, K., X. Zhao. 2011. How dangerous is a day in hospital?: A model of adverse events and length of stay for medical inpatients. *Medical Care*. **49**(12) 1068-1075

Helm, J. E., S. AhmadBeygi, M. P. Van Oyen. 2011. Design and analysis of hospital admission control for operational effectiveness. *Prod. & Oper. Management*. **20**(3) 359-374.

Hoetker, G. 2007. The use of logit and probit models in strategic management research: Critical issues. *Strat. Management J.* **28** 331-343.

Hopp, W. J., S. M. R. Iravani, F. Liu. 2009. Managing white-collar work: An operations-oriented survey. *Prod. & Oper. Management*. **18**(1) 1-32.

Hopp, W. J., S. M. R. Iravani, G. Y. Yuen. 2007. Operations systems with discretionary task completion. *Management Sci.* **53**(1) 61-77.

Hussey, P. S., C. Eibner, M. S. Ridgely, E. A. McGlynn. 2009. Controlling u.S. Health care spending — separating promising from unpromising approaches. *New England J. Medicine*. **361**(22) 2109-2111.

Jenkins, S. P. 2004. *Survival analysis*. Institute for Social and Economic Research, University of Essex, Colchester, UK.

Jouini, O., Y. Dallery, R. Nait-Abdallah. 2008. Analysis of the impact of team-based organizations in call center management. *Management Sci.* **54**(2) 400-414.

KC, D., C. Terwiesch. 2009. Impact of workload on service time and patient safety: An econometric analysis of hospital operations. *Management Sci.* **55**(9) 1486-1498.

KC, D. S., C. Terwiesch. 2012. An econometric analysis of patient flows in the cardiac intensive care unit. *M&SOM*. **14**(1) 50-65.

KC, D. S., C. Terwiesch. 2014. An econometric analysis of emergency patient admission: The role of available beds, active patient discharge, and inpatient demand smoothing. *Working Paper*.

Kellermann, A. L. 2006. Crisis in the emergency department. *New England J. Medicine*. **355**(13) 1300-1303.

Kim, S.-H., C. Chan, M. Olivares, G. Escobar. 2012. Icu admission control: An empirical study of capacity allocation and its implication on patient outcomes. *Available at SSRN 2062518*.

Kuntz, L., R. Mennicken, S. Scholtes. 2013. Stress on the ward: Evidence of safety tipping points in hospitals. *Working Paper, Cambridge Judge Business School, UK*.

Landis, J. R., G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*. **33**(1) 159-174.

Levin, D. C., V. M. Rao. 2008. Turf wars in radiology: Updated evidence on the relationship between self-referral and the overutilization of imaging. *Journal of the American College of Radiology*. **5**(7) 806-810.

Little, J. D. C. 1961. A proof for the queuing formula: $L = \lambda w$. *Oper. Res.* **9**(3) 383-397.

Litvak, E. 2010. *Managing patient flow in hospitals: Strategies and solutions*. Joint Commision Resources, Oakbrook Terrace, IL.

Macario, A., F. Dexter, R. D. Traub. 2001. Hospital profitability per hour of operating room time can vary among surgeons. *Anesthesia & Analgesia*. **93**(3) 669-675.

McManus, M. L., M. C. Long, A. Cooper, J. Mandell, D. M. Berwick, M. Pagano, E. Litvak. 2003. Variability in surgical caseload and access to intensive care services. *Anesthesiology*. **98**(6) 1491-1496.

Meyer, H. 2011. At upmc, improving care processes to serve patients better and cut costs. *Health Affairs*. **30**(3) 400-403.

Netessine, S., R. A. Shumsky. 2002. Introduction to the theory and practice of yield management. *INFORMS Transactions on Education*. **3**(1) 34-44.

Office of Inspector General. 2001. *Medicare hospital prospective payment system: How drg rates are calculated and updated*. http://oig.hhs.gov/oei/reports/oei-09-00-00200.pdf.

Oliva, R., J. D. Sterman. 2001. Cutting corners and working overtime: Quality erosion in the service industry. *Management Sci.* **47**(7) 894-914.

Parkinson, C. N. 1958. *Parkinson's law, or the pursuit of progress*. John Murray.

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

Pines, J. M., J. E. Hollander. 2008. Emergency department crowding is associated with poor care for patients with severe pain. *Annals of Emergency Medicine*. **51**(1) 1-5.

Pitts, S. R., R. W. Niska, J. Xu, C. W. Burt. 2008. National hospital ambulatory medical care survey: 2006 emergency department summary. *Natl Health Stat Report*(7) 1-38.

Powell, A., S. Savin, N. Savva. 2012 Physician workload and hospital reimbursement: Overworked servers generate lower income. *M&SOM*. **14**(4) 512-528.

Powell, S. G., K. L. Schultz. 2004. Throughput in serial lines with state-dependent behavior. *Management Sci.* **50**(8) 1095-1105.

Roberts, R. R., P. W. Frutos, G. G. Ciavarella, L. M. Gussow, E. K. Mensah, L. M. Kampe, H. E. Straus, G. Joseph, R. J. Rydman. 1999. Distribution of variable vs fixed costs of hospital care. *J. Amer. Medical Assoc.* **281**(7) 644-649.

Saghafian, S., W. J. Hopp, M. P. Van Oyen, J. S. Desmond, S. L. Kronick. 2012. Patient streaming as a mechanism for improving responsiveness in emergency departments. *Oper. Res.* **60**(5) 1080-1097.

Schultz, K. L., D. C. Juran, J. W. Boudreau. 1999. The effects of low inventory on the development of productivity norms. *Management Sci.* **45**(12) 1664-1678.

Schultz, K. L., D. C. Juran, J. W. Boudreau, J. O. McClain, L. J. Thomas. 1998. Modeling and worker motivation in jit production systems. *Management Sci.* **44**(12) 1595-1607.

Schultz, K. L., J. O. McClain, L. J. Thomas. 2003. Overcoming the dark side of worker flexibility. *J. Oper. Management*. **21**(1) 81-92.

Sebelius, K. 2013. *Affordable care act at 3: Paying for quality saves health care dollars* Health Affairs Blog.

Shapiro, R. D. 1996. National cranberry cooperative. *Harvard Business School Case*. #**688-122**.

Song, H. A. L. T. K. L. M. 2013. The impact of pooling on throughput time in discretionary work settings: An empirical investigation of emergency department length of stay. *Harvard Business School Working Paper*.

State of California OSHPD Health Information Decision. 2008. *California acute care hospital services statewide trends, 1997-2006*. State of California Office of Statewide Health Planning and Development.

Stidham, S., R. R. Weber. 1989. Monotonic and insensitive optimal policies for control of queues with undiscounted costs. *Oper. Res.* **37**(4) 611-625.

Stiell, I. G., G. A. Wells, K. Vandemheen, A. Laupacis, R. Brison, M. A. Eisenhauer, G. H. Greenberge, I. MacPhail, R. D. Mcknight, M. Reardon, R. Verbeek, J. Worthington, H. Lesiuk. 1997. Variation in ed use of computed tomography for patients with minor head injury. *Annals of Emergency Medicine*. **30**(1) 14-22.

Sun, B. C., J. Adams, E. J. Orav, D. W. Rucker, T. A. Brennan, H. R. Burstin. 2000. Determinants of patient satisfaction and willingness to return with emergency care. *Annals of Emergency Medicine*. **35**(5) 426-434.

Tan, T. F., S. Netessine. 2014. When does the devil make work? An empirical study of the impact of workload on server's performance. *Management Sci.* **Forthcoming**.

Tucker, A. L., S. J. Spear. 2006. Operational failures and interruptions in hospital nursing. *Health Services Res.* **41**(3p1) 643-662.

US National Library of Medicine and National Institutes of Health. 2013. *Medlineplus*.

White, D. L., C. M. Froehle, K. J. Klassen. 2011. The effect of integrated scheduling and capacity policies on clinical efficiency. *Prod. & Oper. Management*. **20**(3) 442-455.

Appendix

Figure A.1. Buildup of **complementary log-log** models for septicemia/severe sepsis w/o MV 96+ hours w/MCC (DRG 871)

Outcome: Hazard at time t, h(t)	(1) *Base*	(2) *Non-busy*	(3) *Admit*	(4) *Discharge*	(5) *Complete*
Day of Week *(Baseline: Sunday)*					
Monday	-0.0471	0.0449	-0.0903	-0.0591	0.0250
Tuesday	-0.0682+	0.2050**	-0.1040*	0.0173	0.1450**
Wednesday	-0.0403	0.2930**	-0.0602	0.0548	0.2150**
Thursday	-0.0216	0.2580**	-0.0345	0.0159	0.1810**
Friday	-0.0191	0.1690**	-0.0292	-0.0230	0.1200**
Saturday	-0.0090	0.0719**	0.0065	0.0050	0.0564*
Year *(Baseline: 2008)*					
2009	0.0577**	0.0383*	0.0527*	0.0368*	0.0288*
Race *(Baseline: Blank)*					
White	-0.213*	-0.158*	-0.226*	-0.160*	-0.153*
Black	-0.280*	-0.142+	-0.299**	-0.175*	-0.157*
Hispanic	-0.225*	-0.192*	-0.235*	-0.195*	-0.185*
Asian/Pacific Islander	-0.349**	-0.240**	-0.358**	-0.249**	-0.238**
Native American/Eskimo/Aleut	0.1620	-0.1280	0.1610	-0.1530	-0.1780
Other	-0.303*	-0.207*	-0.325**	-0.203*	-0.198*
Age					
Age	0.0155+	0.0118+	0.0154+	0.0100	0.0109+
Age^2	-0.0001*	-8.89e-05*	-0.0001*	-7.49e-05+	-8.16e-05*
Gender *(Baseline: Male)*					
Female	0.0501*	0.0026	0.0493*	-0.0073	-0.0083
Other	-0.0587	-0.638**	-0.108+	-0.267**	-0.280**
Insurance Type *(Baseline: Blank)*					
Medicare	-0.935**	-0.3360	-1.007**	-0.415**	-0.433**
Medi-Cal	-1.420**	-0.562*	-1.493**	-0.631**	-0.645**
Private Coverage	-0.836**	-0.2620	-0.910**	-0.343*	-0.360**
Workers' Compensation	-1.108**	-0.2830	-1.180**	-0.367*	-0.389**
County Indigent Programs	-1.393**	-0.3000	-1.466**	-0.369*	-0.390**
Other Government	-1.110**	-0.3240	-1.195**	-0.403*	-0.412**
Other Indigent	-1.225*	-0.2420	-1.310**	-0.414+	-0.385+
Self Pay	-1.051**	-0.3300	-1.127**	-0.422*	-0.441**
Other Pay	-1.005**	-0.3510	-1.088**	-0.364+	-0.391*
Occupancy on Day of Admission					
Medical					
Main Effect	-	-0.0107**	-0.0003	-	-0.0099**
Busy (≥85% & <93%)	-	-	-0.0041	-	0.0099**
Very Busy (≥93%)	-	-	0.0166	-	-0.0013
Surgical					
Main Effect	-	-0.0065**	-0.0015	-	-0.0053**
Busy (≥85% & <93%)	-	-	-0.0059	-	0.0005
Very Busy (≥93%)	-	-	0.0177	-	0.0044
Obstetrics					
Main Effect	-	1.09e-05	0.0020**	-	0.0008
Busy (≥70% & <85%)	-	-	-0.0080**	-	-0.0021
Very Busy (≥85%)	-	-	0.0105+	-	-0.0006
Occupancy on Day of Discharge					
Medical					
Main Effect	-	0.0206**	-	0.0225**	0.0245**

Busy (≥85% & <93%)	-	-	-	-0.0573**	-0.0539**
Very Busy (≥93%)	-	-	-	0.0495**	0.0450**
(Figure A.1. Continued)					
Surgical					
Main Effect	-	0.0157**	-	0.0153**	0.0160**
Busy (≥85% & <93%)	-	-	-	-0.0400**	-0.0351**
Very Busy (≥93%)	-	-	-	0.0420**	0.0365**
Obstetrics					
Main Effect	-	0.0054**	-	0.0080**	0.0072**
Busy (≥70% & <85%)	-	-	-	-0.0128**	-0.0114**
Very Busy (≥85%)	-	-	-	0.0048	0.0036
Admissions on Day of Admission					
Medical					
Main Effect	-	-0.0030**	0.0007	-	-0.0025**
Busy (≥70% & <85%)	-	-	-0.0009	-	0.0009
Very Busy (≥85%)	-	-	0.0011	-	0.0006
Surgical					
Main Effect	-	-0.0008+	0.0013	-	-0.0008
Busy (≥70% & <85%)	-	-	-0.0049+	-	-0.0015
Very Busy (≥85%)	-	-	0.0130*	-	0.0068*
Obstetrics					
Main Effect	-	-0.0005+	-0.0003	-	-0.0009+
Busy (≥60% & <80%)	-	-	-0.0020	-	-0.0001
Very Busy (≥80%)	-	-	0.0083+	-	0.0055+
Admissions on Day of Discharge					
Medical					
Main Effect	-	0.00574**	-	0.0078**	0.0073**
Busy (≥70% & <85%)	-	-	-	-0.0170**	-0.015**
Very Busy (≥85%)	-	-	-	0.0152**	0.0135**
Surgical					
Main Effect	-	0.0001	-	0.0012*	0.0008+
Busy (≥70% & <85%)	-	-	-	-0.0070**	-0.0063**
Very Busy (≥85%)	-	-	-	0.0091+	0.0095+
Obstetrics					
Main Effect	-	0.0012**	-	0.0035**	0.0030**
Busy (≥60% & <80%)	-	-	-	-0.0102**	-0.0089**
Very Busy (≥80%)	-	-	-	0.0107**	0.0098**
Day of Stay					
Day 3	0.573**	1.031**	0.565**	1.040**	1.040**
Day 4	1.861**	2.118**	1.854**	2.117**	2.115**
Day 5	2.171**	2.405**	2.167**	2.396**	2.397**
Day 6	2.212**	2.422**	2.210**	2.415**	2.416**
Day 7	2.223**	2.470**	2.221**	2.490**	2.489**
Day 8	2.062**	2.322**	2.062**	2.345**	2.344**
Day 9	1.438**	1.588**	1.438**	1.595**	1.599**
Disposition			Included (Not Significant)		
Comorbidities/Elixhauser Index			Included		
Hospital			Included		
Constant	-1.443**	-4.146**	-1.379*	-5.732**	-4.611**
Observations (patient-days)	46,117	46,102	46,102	46,102	46,102

**P<0.01, *P<0.05, +P<0.1

Figure A.2. Confirmation that patient characteristics did not change pre-/post-change

	Flagship		Community	
	Pre Change	***Post Change***	***Pre Change***	***Post Change***
Age (Mean)	40.8	41.1	44.2	45.1
	*T-test**	*p=0.824*	*T-test**	*p=0.5182*
Sex				
Female	137	3,360	124	2,723
Male	87	1,886	68	1,385
	Chi-square	*p=0.378*	*Chi-square*	*p=0.626*
Race				
White	148	3,168	139	2,917
Black	30	895	27	643
Other	46	1,183	26	548
	Fisher's exact	*p=0.209*	*Fisher's exact*	*p=0.861*
ESI				
1	0	6	0	1
2	58	1,186	6	177
3	163	3,783	182	3,794
4	3	96	4	94
5	0	7	0	8
	Fisher's exact	*p=0.811*	*Fisher's exact*	*p=0.849*
Day of Week				
Sunday	37	1,186	44	963
Monday	61	1,524	58	1,102
Tuesday	70	1,448	63	1,044
Wednesday	70	1,399	51	983
Thursday	75	1,328	59	961
Friday	67	1,342	39	961
Saturday	49	1,177	41	936
	Chi-square	*p=0.082*	*Chi-square*	*p=0.317*
Visit times				
Weekday	205	4,158	163	2,842
Night/Weekend	224	5,246	192	4,108
	Chi-square	*p=0.145*	*Chi-square*	*0.061*

Comparison of patient characteristics before and after processing change. All but Day of Week and Visit times were run for night/weekend patients only (Day of Week and Visit times run on all patients at each hospital), but similar results were obtained when using all patients. Other than Age, each analysis compares number of patients in each category. *Also ran using Wilcoxon-Mann-Whitley test, with similar results.

Figure A.3: Categories and medical indications, as described by the American College of Emergency Physicians, for which an emergency U/S is warranted

U/S Category *(Column 1)*	**Medical Indications** *(Column 2)*
Abdominal Aortic Aneurysm (AAA)	AAA
Biliary	Cholelithiasis
	Cholecystitis
	Common bile duct abnormalities
	Liver abnormalities
	Portal vein abnormalities
	Abnormalities of the pancreas
	Other gallbladder abnormalities
	Unexplained jaundice, ascites
Echocardiography	Pericardial effusion and/or tamponade LV systolic function
	RV function and/or acute pulmonary hypertension w/unexplained chest pain
	Dyspnea or hemodynamic instability
Pelvic	Intrauterine/ectopic pregnancy
	Ovarian cysts
	Fibroids
	Tobu-ovarian abscess
Renal	Obstructive uropathy and/or urinary retention
	Acute hematuria
	Renal failure
	Infection/abscesses
	Bladder and prostate abnormalities
Trauma	Fluid in peritoneal, pericardial, and pleural cavities
	Pneumothorax
	Solid organ injury
Venous Thrombosis	Acute proximal DVT in lower extremities
	Chronic DVT
	Distal DVT
	Superficial venous thrombosis
	Lower extremity swelling/pain
	Cellulitis
	Abscess
	Muscle hematoma
	Fasciitis
	Baker's cyst
	Upper extremity venous thrombosis

Figure A.4: Symptoms associated with U/S use

Abdominal Cramping	Itching
Abdominal Pain/Pressure	Joint Pain/Swelling/Stiffness
Abdominal Swelling (ascites)	Leg Pain
Abnormal films/CT	Lump (mass) in abdomen
Abnormal Vaginal Bleeding	Malaise/Not Feeling Well
Alcohol/Drug	Muscle aches and pains/Body Aches
Anxiety/nervousness	Nausea
Back pain/cramps	Neck Pain
Blood in Stool/Abnormal Stool Color or smell	Numbness
Blood in Urine	Other GI History
Chest Pain/Tightness	Pain or Burning with urination
Chills	Pelvic Pain
Confusion/Altered Mental Status/Unresponsive	Post Choley/Gallbladder removal
Cough	Post Surgery
Crohn's Disease	Post-op renal
Decreased amount of urine	Rule out (R/o) AAA
Diarrhea	R/o Biliary
Difficulty breathing/SOB	R/o DVT or U/S guided procedure
Diminished Appetite	R/o DVT/US guided procedure- abscess
Dizziness	R/o echo
Double or Blurred Vision	R/o pelvic
Easy Bruising	R/o renal
Fainting/LOC	R/o trauma
Fast/Rapid HR/ Palpitations/Pounding Heartbeat	Rib Pain
Fatigue	RUQ Abdominal Pain
Feeling Faint	Seizure
Fever	Shakiness
Flank Pain	Shoulder Pain
Foot Pain/Swelling	Skin sore or rash
Frequent Urge to Urinate	Small Bowel Obstruction/Constipation/BM Pain
Gaseous	Sweating
Groin Pain	Swelling of extremities
Headache	Tube issues
Hernia	Vomiting
History of liver, pancreas, gallbladder issues	Vomiting Blood
History of renal/Post-op	Warm tissue
Hyper- or hypoglycemia	Weakness
Hypertension	Weight gain
Hypotension	Weight loss
Inability to urinate	Yellowing of skin (Jaundice)