University of Massachusetts Amherst

Stochastic Models for Capacity Planning in Healthcare Delivery: Case Studies in an Outpatient, Inpatient and Surgical Setting

Item Type	dissertation
Authors	Ozen, Asli
DOI	10.7275/ybd9-1z08
Download date	2025-04-09 08:57:41
Link to Item	https://hdl.handle.net/20.500.14394/17447

STOCHASTIC MODELS FOR CAPACITY PLANNING IN HEALTHCARE DELIVERY CASE STUDIES IN AN OUTPATIENT, INPATIENT AND SURGICAL SETTING

A Dissertation Presented

by

ASLI ÖZEN

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2014

Mechanical and Industrial Engineering

© Copyright by Asli Özen 2014

All Rights Reserved

STOCHASTIC MODELS FOR CAPACITY PLANNING IN HEALTHCARE DELIVERY CASE STUDIES IN AN OUTPATIENT, INPATIENT AND SURGICAL SETTING

A Dissertation Presented

by

ASLI ÖZEN

Approved as to style and content by:

Hari Balasubramanian, Chair

Ana Muriel, Member

Senay Solak, Member

Joan Roche, Member

Donald Fisher, Department Chair
Mechanical and Industrial Engineering

This dissertation is dedicated to my mother who has been my role model and a constant source of inspiration.

ACKNOWLEDGEMENTS

Though only my name appears on the cover of this dissertation, a great many people have contributed to its production. I owe my gratitude to all those people who have made this dissertation possible and because of whom my graduate experience has been one that I will cherish forever.

I owe in large measure the stimulation of my thoughts to my adviser and mentor Professor Hari Balasubramanian. His unique vision allowed me to find a common ground between theory and applied research, and pursue this dissertation. I am forever indebted to him for his generous support, encouragement and guidance. I could not have imagined having a better advisor and mentor for my Ph.D. study. And, I thank my dissertation committee, Professor Ana Muriel, Professor Joan Roche and Professor Senay Solak for their encouragement, insightful comments, helpful career advice and suggestions in general.

I would like to thank my family for all their unconditional love, support and encouragement. For my parents, Seza and Haluk, who raised me with a love of science and supported me in all my pursuits. I am deeply grateful to them for being the best parents imaginable and, to my brother, Can, for providing me with a compass every time I need one. This Ph.D. is a testament to your faith in me, I

hope I have made you proud. To my close friend Baris who has made the completion of this degree and dissertation bearable and enjoyable; to my loyal friend Duygu for the meaning she adds to my life; to my best friend Ceyda for brightening up my days with her joy; to Joanne and Michael who has provided me with friendship, support, encouragement, advice, and laughter as needed through the many ups and downs that have accompanied the completion of this degree and dissertation; to Catherine for helping me redefine what family is; to Jivan who has reminded me to believe in the universe for everything happens for a reason, everytime I needed to be reminded of; to Ilke who has been the big-sister I never had and for making me a better person; and to Jose, Jocelyn, Berra, Piril, Melih and all my beloved friends in the Valley for helping me sustain my determination and resilience through rough times. Thank you; this dissertation could not have been completed without your loving support.

I was recruited at UMass as part of the Hluchyj Engineering-Nursing fellowship program on September 2010, aimed at improving interdisciplinary collaboration between engineers and healthcare providers. We contacted Joan Roche on November 2010 to start working on an interdisciplinary project, which led to our first meeting at Baystate Medical Center (BMC) with Patty Samra on December 2010. It took us from January to May 2011 to define the problem and data requirements with our multidisciplanary team, composed of me, my advisor Hari Balasubramanian, Patty Samra, Joan Roche, Haiping Li, Mike Ehresman and Todd Fairman. The results and the model were presented to stakeholders on Summer 2012. The stakeholders include: ED, Surgical, Finance, Nursing, BMC Leadership and Process improvement Groups. BMC provided me with seed funding ($13,500) to continue our project on

improving the patient flow on November 2012. I am grateful for the generosity of Terry and Mike Hluchyj who have provided the Hluchyj fellowship that both enabled and initiated the research on inpatient flow problem. I would like to thank our team at BMC that has been a great source of inspiration with their invaluable comments and contributions at every stage of the project.

I would like to acknowledge all of my team members in Mayo Clinic for their efforts and contributions in carrying out our research in surgical care. The research team included my advisor Hari Balasubramanian; Yariv Marmor, who was previously a post-doc in Health Sciences Research Department in Mayo Clinic and is now at La Braude University, Israel; Tom Rohleder, who used to be Professor of Healthcare Systems Engineering in Health Services Research Department; Paul Huddleston the spine surgeon in Mayo Clinic who has initiated the project and Jeanne Huddleston who has provided invaluable insights with her background as both being an internist and the director of the Health Services Research Department in Mayo Clinic. Also I am grateful for the Center for Excellence who was responsible for the development of the web-based interface that led to the implementation of the pilot study.

This work came to life through the funding in part by the Agency of Healthcare Research and Quality (AHRQ) Grant R03 HS18795-01. I am very grateful for generous resources provided by the Department of Mechanical and Industrial Engineering at UMass, Amherst and to our inspirational department chair Professor Fisher.

ABSTRACT

STOCHASTIC MODELS FOR CAPACITY PLANNING IN HEALTHCARE DELIVERY CASE STUDIES IN AN OUTPATIENT, INPATIENT AND SURGICAL SETTING

MAY 2014

ASLI ÖZEN

B.Sc., BILKENT UNIVERSITY

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Hari Balasubramanian

The U.S. healthcare system has become far too complex and costly to sustain. As Green points out in her M&SOM editorial, there has never been a more opportune time for operations research to provide guidelines on medical decision making and improving the healthcare delivery process (Green [2012]). We study capacity planning in healthcare while considering the case-mix of patients, using stochastic modeling in different application areas: primary care, inpatient bed allocation and

(spine) surgery scheduling. The research questions we have addressed are relevant and may be of interest to many researchers and practitioners.

Primary Care: The passage of Affordable Care Act (ACA) and the significant influx of insured individuals create an urgent need to increase the effective primary care capacity. Our research in primary care provides a tool to assess supply-demand dynamics, conduct capacity planning and practice design for provider teams. The main objective of Chapter 2 is to optimize the patient mix of primary care physicians in a group practice in order to maximize patient-clinician continuity and access. We use an optimization in a newsvendor-like framework and propose simple, yet near-optimal heuristics. To model case-mix, we use the number of simultaneous chronic conditions (count of comorbidities) a patient has as a predictor of the number of appointment requests. In Chapter 3, we extend this work and use queuing theory to develop methodologies to quantify and evaluate access to care and continuity of care for patient visits with different urgencies. We find that case-mix is a crucial factor to consider in primary care practice design. Further, both panel redesign and capacity pooling can be effective strategies for primary care practice improvement. In particular, even a little capacity pooling can make a big difference.

Inpatient Care: Inability to satisfy the bed requests in a timely fashion for admitted patients leads to emergency department (ED) crowding, ambulance diversions, patients left without being seen, post-anesthesia care unit (PACU) holds and delays, surgery cancelations, and overall in decline in care and safety. One of the major contributing factors to this patient flow gridlock is delayed discharges. We develop an empirically calibrated simulation model to represent a time-varying multi-server

queuing network model with multiple patient classes in Chapter 4. This model is used as a decision support mechanism for inpatient bed planning at Baystate Medical Center, Springfield MA. Our main focus has been on quantifying the impact of discharge profiles to alleviate inpatient bed congestions. A discharge profile is defined by (a) discharge window, which specifies which hours of the day discharges are allowed; and (b) the maximum capacity for discharges in each hour of the window. We conclude that a more responsive discharge policy that prioritizes discharges in units with longer admission queues can significantly reduce waiting times (40% reduction in queue size). On the other hand, an early-in-the-day discharge policy has limited impact on improving bed congestions; we also find that early in day discharges are very hard to implement in practice.

Surgical Care: Due to the length and variability of spine surgeries (Dexter et al. [2010])scheduling is a difficult and important aspect to patient access, effective operations, and financial performance for the spine surgery practice. The main objective of our research in Chapter 5 is to create better patient access and improve revenue as a result of increased surgical capacity with more efficient schedules and an improved patient mix. A multi-stage mixed integer optimization framework has been developed into a web-based application to be used in a pilot study that allowed the surgeons and schedulers to interactively identify best surgical days with patients. A pilot implementation resulted in a utilization increase of 19% and a reduction in overtime by 10%.

This body of work was developed over four years of collaborative research with hospitals and healthcare providers. To ensure that the models are clinically relevant,

we have collaborated extensively with healthcare stakeholders: the spine surgery team at Mayo Clinic and the nursing group at Baystate Medical Center. Our main objective has been to develop data-driven analytical tools for managing networks of healthcare resources to smooth workloads and improve access to care.

TABLE OF CONTENTS

Page

ACKNOWLEDGEMENTS . . . . . **v**

ABSTRACT . . . . . **viii**

LIST OF TABLES . . . . . **xviii**

LIST OF FIGURES . . . . . **xxi**

CHAPTER

1. INTRODUCTION **1**

1.1 Industrial Engineering in Healthcare 1
1.2 Research Motivation 3

1.2.1 Primary care 3
1.2.2 Inpatient care 5
1.2.3 Surgical care 6
1.2.4 Salient features 7

1.3 Methodology 8

1.3.1 Mixed integer programming 8
1.3.2 Stochastic modeling 9

1.3.2.1 Queuing theory and modeling 9
1.3.2.2 Newsvendor model 10

1.3.2.3 Simulation (discrete event and object-oriented simulation) . . . 11

1.4 Dissertation Overview . . . 12

2. THE IMPACT OF CASE-MIX ON TIMELY ACCESS TO APPOINTMENTS IN A PRIMARY CARE GROUP PRACTICE . . . 13

2.1 Introduction . . . 13
2.2 Literature Review . . . 16
2.3 Patient Classification . . . 19
2.4 Example of 4 Physicians . . . 21
2.5 Feasibility of Panel Redesign . . . 23
2.6 The Panel Redesign Formulation (PRF) and Analytical Results . . . 26

2.6.1 The unequal capacity case . . . 32
2.6.2 Deriving the reference overflow O_{ref} . . . 33
2.6.3 $O_{ref} - O_{opt}$ for common cases in practice . . . 36
2.6.4 Summary of contributions . . . 39

2.7 Heuristics . . . 40
2.8 Case Study . . . 42

2.8.1 Data description . . . 42
2.8.2 Panel redesign for test practices . . . 45
2.8.3 Quantifying the price of continuity . . . 51
2.8.4 Impact on other measures . . . 53

2.9 Conclusions and Implications for Practice . . . 57

3. PRIMARY CARE PRACTICE DESIGN UNDER CASE-MIX: JOINT CONSIDERATION OF ACCESS TO CARE AND CONTINUITY OF CARE . . . 61

3.1 Introduction . . . 61
3.2 Methods . . . 66

3.2.1 The models . . . 66

3.2.2 Data and model parameters 70
3.2.3 Model analysis 71

3.3 Results 73

3.3.1 Impact of case-mix on provider utilization 73
3.3.2 Comparison of practice designs under different case-mixes 75

3.4 Discussion 80

4. MODELING HOSPITAL-WIDE PATIENT FLOWS USING SIMULATION 84

4.1 Introduction 84
4.2 Discharge Planning 91
4.3 Literature Review 94

4.3.1 Hospital-wide flow models 94
4.3.2 Why simulation and not queuing? 96

4.4 Data Collection and Analysis 101
4.5 Simulation Model and Analysis 107

4.5.1 Replications 109

4.6 Analyzing the Impact of Discharge Policies 110

4.6.1 Baseline 110
4.6.2 **DP2:** Maximum capacity of 10 in each hour from 10 AM-7 PM, no prioritization 111
4.6.3 **DP3:** Early in the day discharge policy, 10 AM-7 PM, no prioritization 111
4.6.4 Expanded discharge windows 112
4.6.5 Prioritization of discharges 114
4.6.6 **DP9:** 24-hour discharge 115

4.7 Results of the Simulation 115

4.7.1 Validation 115
4.7.2 Impact of discharge policies 118

4.7.2.1 Unit specific analysis 122
4.7.2.2 Waiting time analysis 123
4.7.2.3 Resulting discharge capacities 123

4.8 Discussion and Conclusion 128
4.9 Ongoing Research 131
4.10 Incorporating Transfer Activities Among Different Units 132
4.11 Hospitalist Scheduling Problem 134

4.11.1 Literature review 134
4.11.2 The problem 137
4.11.3 Observations from BMC 138

5. OPTIMIZING SPINE OR SURGICAL THROUGHPUT: ENGINEERING A PULL SYSTEM FOR OUTPATIENT ACCESS 141

5.1 Introduction 141

5.1.1 Historical spine surgery scheduling at Mayo Clinic 142
5.1.2 Objectives 145
5.1.3 Approach 145

5.2 Literature Review 146

5.2.1 Contributions 148

5.3 Optimization Models 150

5.3.1 First stage IP optimization 151
5.3.2 Second stage IP optimization 155
5.3.3 Third stage IP optimization 159

5.4 Case Study 162

5.4.1 Data and model assumptions 162

5.4.1.1 Surgery type: 163
5.4.1.2 Multiple days staged surgery (MDSS) patients: 164
5.4.1.3 Financial analysis and length of stay (LOS): 165

5.4.2 Simulation for scenario generation 167

5.4.2.1 Surgery steps and times: 167
5.4.2.2 Why did we use a simulation model for outcomes projection? 170
5.4.2.3 Parameters and scenarios: 170
5.4.2.4 Inputs to the optimization model: 172

5.4.3 Example optimization results 172

5.4.3.1 First stage optimization: 173
5.4.3.2 Second stage optimization: 174
5.4.3.3 Third stage optimization: 175
5.4.3.4 Simulation for testing the robustness: 176
5.4.3.5 Sensitivity analysis 176

5.5 Implementation 178

5.5.1 Results of the pilot implementation 179
5.5.2 Lessons learned from the pilot 183

5.6 Conclusions 185

6. FUTURE WORK 188

6.1 Opportunities in Primary Care 188

6.1.1 Testing the applicability of the findings for primary care on a national level 188
6.1.2 The Patient Centered Medical Home (PCMH) in primary care 189
6.1.3 Studying the relationship between readmissions and access to primary care 190

6.2 Opportunities in Inpatient Bed Planning 192

6.2.1 Pre and post allocation delays 192
6.2.2 Modeling readmissions based on different discharge policies 193
6.2.3 Incorporating uncertainty to the discharge process 193

6.3 Opportunities in Spine Surgery Scheduling 194

6.3.1 Modeling other types of uncertainties 194
6.3.2 Extended surgery scheduling model in the presence of other resources and uncertainty 194
6.3.3 Testing the robustness of the model by extensions to other surgical services 195

APPENDICES

A. QUEUING MODEL FORMULAS FOR CHAPTER 3 196
B. ADDITIONAL DATA ANALYSIS ON INPATIENT CARE 202
C. ALGORITHM OF THE INPATIENT FLOW SIMULATION MODEL 210
D. ARENA MODEL 217
E. INITIAL ANALYSIS ON ONGOING WORK IN INPATIENT CARE 222
F. ADDITIONAL DATA ANALYSIS ON SURGICAL CARE 225
G. SENSITIVITY ANALYSIS 231

BIBLIOGRAPHY 240

LIST OF TABLES

Table **Page**

2.1 Mean and standard deviation of visits in 2006, for patients with different counts of comorbidities 21

2.2 Case-mix, panel size and performance measures for 4 physicians 23

2.3 Binomial p_i values for each patient category 43

2.4 Results for Test Practice 1 46

2.6 Results for Test Practice 3 49

2.7 Results for Test Practice 4 50

2.8 Price of continuity in terms of number of patients, where TP: Test Practice 51

2.5 Results for Test Practice 2 56

3.1 Arrival rate per patient per day for each category. 74

3.2 Example of 7 hypothetical panels 74

3.3 Case-mixes of Physicians 1 and 2: Initial/Baseline panels 75

3.4 Case-mixes of Physicians 1 and 2: Balanced Panels/After redesign 75

3.5 Design comparison under the baseline and balanced panels 78

4.1 Earlier in the day discharge 88

4.2 Later in the day discharge 89

4.3 Unit specific analysis 107

4.4 Connecting letters report for queue size 119

4.5 Average queue size 122

4.6 Queue size percentiles using different discharge policies 123

4.7 Time motion study 140

5.1 Properties of multi-segment surgeries 165

5.2 Distributions for the time-stamps 169

5.3 Simulation inputs to the optimization model 173

5.4 Optimal 12 week blue-orange schedule 175

5.5 Pre and post-implementation results for all surgeons 182

B.1 Percentage of "APR-DRG Severity of Illness" categories by days of the week 209

D.1 Results of the Arena model 221

E.1 Impact of restricting the number of discharges on queue size 223

E.2 Impact of transfers on queue size 224

F.1 Patient characteristics 226

F.2 Clinical characteristics 226

G.1 Parameter estimates for access 236

G.2 Parameter estimates for NOI 237

G.3 Parameter estimates for utilization 238

LIST OF FIGURES

Figure **Page**

2.1 A visual summary of the panel redesign problem to minimize the maximum overflow . . . 30

2.2 Comparison of O_{ref} and O_{opt} as a function of Z_{CP} for the 2-physician and 4-physician example . . . 38

2.3 Results for 2 physicians with equal capacity . . . 54

3.1 Practice designs . . . 69

3.2 The impact of partial pooling on access to care and continuity of care for both baseline and balanced panels . . . 79

4.1 Admission and discharge rates . . . 86

4.2 Queuing framework . . . 101

4.3 Admission and discharge process . . . 102

4.4 Arrival pattern by patient sources . . . 104

4.5 Volume and LOS values of patient sources . . . 105

4.9 Capacity thresholds and actual realizations . . . 124

4.6 Queue size ANOVA analysis . . . 126

4.7 Waiting time ANOVA analysis 127

4.8 Discharge profile for two units Medical Telemetry and Surgical/Orthopedic 127

4.10 Best solution 129

5.1 Cumulative distribution of the surgery time by each patient category 164

5.2 Relationship between NOI and LOS for Medicare and Non-Medicare patients 166

5.3 The percentage of weekend overflow by each patient category and surgery DOW 168

5.4 Stages in OR time 169

5.5 Percentage of overtime and utilization by patient category 171

5.6 Validation of simulation model with 95th percentile 172

5.7 Optimal surgery mix 174

5.8 Trade-off between utilization and NOI 177

5.9 SSSO screenshots 179

5.10 Question screen to categorize surgical case 180

5.11 Initial screen that identifies optimal days 180

5.12 Comparison of output measures evaluated during the pilot 181

B.1 ALOS and volume for each MDC 203

B.2 Arrivals by hour – 50th percentile 204

B.3 Arrival rate of ED patients on each DOW and hours of the day 205

B.4 Arrival rate of elective surgeries 206

B.5 Distribution of the LOS (hours) for patients from MDC 4 207

B.6 Distribution of the LOS for patients from MDC 22 208

B.7 LOS and volume of elective patients presented over days of the week 209

D.1 Illustration of Arena model 220

F.1 The BOD distribution 227

F.2 OR cleaning time distribution 229

F.3 Calculation of OR cleaning time 230

G.1 Partitioning for access 233

G.2 Partitioning for NOI 234

G.3 Partitioning for utilization 235

G.4 Regression analysis for access 237

G.5 Regression analysis for NOI 238

G.6 Regression analysis for utilization 239

CHAPTER 1

INTRODUCTION

1.1 Industrial Engineering in Healthcare

Effective and efficient delivery of healthcare has become a major concern in the U.S. With over $2.3 trillion/year spent on U.S. healthcare, it is one of the most expensive health systems in the world. Nearly 15% of the GDP has been spent on healthcare in the U.S. (much higher than developed nations' average), with an unsatisfied patient population both in terms of quality and access (Mahon and Weymouth [2012]). The aging population, increase in chronic conditions and significant influx of new patients covered– 32 million more people who will have insurance by 2019 (Manchikanti et al. [2011]) under the Affordable Care Act (ACA)– will create an even bigger discrepancy between supply and demand (Schoen et al. [2011]).

In order to improve the healthcare delivery process, operations research has been applied to healthcare systems since 1916 (Gilbreth [1916]). There were more than a hundred publications on analytical models applied to healthcare, as early as 1980s (Fries [1980]). However,this field has been experiencing a renaissance in the last 10 years. The abundance of data has triggered this growth in health systems en-

gineering as well. As Linda Green points out in her M&SOM editorial, there has never been a more opportune time for operations research to provide guidelines on medical decision making and improving the healthcare delivery process (Green [2012]). "Numerous studies agree that roughly 30% of total U.S. healthcare costs are attributable to inefficient poorly designed processes, prompting publications by the National Academy of Engineering (NAE) and Institute of Medicine (IOM) to advocate much greater application of systems engineering, operations research, management science" (IOM and NAE [2012], Grossmann et al. [2011]). In their joint report IOM and NAE discuss an action plan on how to integrate systems engineering to healthcare delivery in order to achieve the six "quality aims" set by IOM, which are safe, effective, timely, patient-centered, efficient and equitable healthcare system (IOM et al. [2005]). What's more important is that policymakers have come to realize the need for analytical skills in designing healthcare delivery.

The areas for operations research applications have diversified, from medical decision making (Sox et al. [2013]) to lean management in healthcare (Kollberg et al. [2006]). In her editorial Green points out that operations research can address a wide spectrum of problems from macro to micro level decisions. Macro or strategic decisions involve policy level decisions that is related to the supply of major healthcare resources like hospital beds; whereas, micro or operational level deals with decisions made on a daily basis, most commonly related to issues of process design and resource allocation (Green [2012]). This dissertation will focus on operational level and not on macro level decisions.

From an operational perspective, operations research methods can help managers plan and manage capacity to meet wait time targets (Patrick and Puterman [2008]). We study capacity planning in healthcare while considering the case-mix of patients, using stochastic modeling in different application areas: primary care, inpatient bed allocation and (spine) surgery scheduling. In what follows, I will provide the motivating reasons for our research in primary, inpatient and surgical care, and summarize the focus of this dissertation.

1.2 Research Motivation

1.2.1 Primary care

With the recent passage of the Patient Protection and Affordable Care Act, the uninsured population is expected to decrease by more than half. However, many areas in U.S. are already facing severe shortage in primary care workforce. According to the Commonwealth Fund, Americans are reporting greater difficulty in achieving timely urgent appointments, not including emergency departments (EDs), compared to the other developed countries (Schoen et al. [2011]). Estimates for the capacity of primary care physicians (PCPs) report one physician for every 2500 patient, which is an unsustainable number for the continuity of care requirements (Alexander et al. [2005]). Hofer et al. [2011] show that 15 to 24 million more primary care visits are expected as a result of the increase in demand from ACA. And this translates to an additional 6000 PCPs required to accommodate the increase in demand.

Compounding the increase in patient volumes and the shortage of primary care workforce is the aging population and the epidemic of chronic diseases, which will

likely give rise to more patients with multiple comorbidities, requiring more physician time and resources. Currently, "45% of the U.S. population have chronic conditions requiring care management. Of this population, 60 million, or roughly half of those with chronic conditions, have multiple conditions" (Kopach-Konrad et al. [2007]). From a financial perspective, chronic disease management account to 75% of the whole medical spending (CDC [2011]). So improving access to PCPs is crucial for better health outcomes as well as decreasing medical costs.

In order to address this new influx of patients, many practices are engaged in transforming into Patient Centered Medical Homes (PCPCC [2013]). The main objective of medical homes is to form a coordinated and integrated care team that provides patient centered care. Yet, there is a lack of analytical methods that can inform the formation of such teams and the allocation of workload among different team members to achieve the best outcome. Our primary care study provides a tool to assess the supply demand dynamics, conduct capacity planning and practice design for primary care teams.

We formulate the problem of minimizing the maximum overflow (probability that the demand will exceed the capacity) for a multi-physician practice as a non-linear integer programming problem and establish structural insights that enable us to create simple yet near optimal heuristic strategies to change panels (set of patients the physician is responsible from). This optimization framework helps a practice: 1) quantify the imbalances across physicians due to the variation in case-mix and panel size, and the resulting effect on access; and 2) determine how panels can be altered in the least disruptive way to improve access.

1.2.2 Inpatient care

Inability to satisfy the bed requests in a timely fashion for inpatients leads to ED crowding, ambulance diversions, patients left without being seen, post-anesthesia care unit (PACU) holds, operating room (OR) delays, surgery cancellations, and overall decline in care and safety (Green [2003]). One of the major contributing factors to this patient flow gridlock is delayed discharges. Late discharges are typically the result of the timing of physician rounds, lack of coordination with the patients' family members about the discharge time and delays resulting from post-acute care facilities.

It is essential to identify capacity levels for hospital beds in conjunction with finding admissions and discharge policies that will be the most cost effective (Green [2012]). And our goal with our nursing collaborators in Baystate Medical Center (BMC) is to provide guidelines on how hospitals should manage their discharge capacity in the presence of demand, LOS and discharge variability. This research enabled us to develop insights to reduce waiting times for inpatient beds from all patient sources. We use an empirically calibrated discrete event simulation to quantify the impact of discharge timing on timely access to inpatient beds. We evaluate both quantitatively and qualitatively, various discharge policies including expanding discharge windows, limiting the number of discharges to a threshold and prioritizing discharges based on the admissions queue. In particular, a more responsive discharge policy that prioritizes discharges for those units that have the longest admission queues (prioritization scheme) results in significant improvement in decreasing waiting times.

1.2.3 Surgical care

As pointed out in Nan and Li [2011], surgical suites management impacts costs, patient flow and resource utilization throughout the whole hospital. And especially for spine surgeries, due to the length and variability, scheduling is a difficult and important aspect to patient access, effective operations, and financial performance (Dexter et al. [2010]). Further complicating factors for scheduling and OR management in Mayo Clinic Spine Practice are emergency cases, short-term cancellations, and complex cases that require more than one surgery to address a patient's medical needs. The primary objective of our research is to create better patient access as a result of increased surgical capacity with more efficient schedules. We not only maximize surgeon and OR utilization but also incorporate profitability while keeping overtime and potentially unsafe surgical days under control. Our model was implemented at Mayo Clinic in a controlled pilot and we evaluate the results of the intervention. A pilot implementation resulted in an increase in utilization of 19%, a reduction in overtime by 10% and an increase in average NOI per case by 22%. In summary, the pilot implementation was deemed successful, but not as comprehensively as desired.

All of these bodies of work were developed over four years of collaborative research with hospitals and healthcare providers. To ensure that the models are clinically relevant, we have collaborated extensively with various healthcare stakeholders: the spine surgery team at Mayo Clinic and the nursing group at BMC. Our main objective in this dissertation is to develop data-driven analytical tools for managing workloads in networks of healthcare resources to smooth workloads and improve access to care.

1.2.4 Salient features

Clearly different areas of healthcare have different problems and are in need of different solutions. However, there are some common distinguishing features present in all of these problems in this dissertation.

(1) Firstly we observe a heterogeneous demand which requires us to model the case-mix of patients using data mining. For instance, we categorize patients based on their comorbidity counts in primary care, patients' major diagnostic categories (MDC) and admission source in inpatient setting and clinical characteristics in surgery scheduling. These categorizations are both clinically and statistically relevant.

(2) As in most healthcare applications the objective is generally not only financially oriented, but rather a multi-objective function, which aims to maximize patient satisfaction and access to care as well.

(3) Our key methods are almost always data driven modeling. We use data from Mayo Clinic Rochester, MN (both from Primary Care Internal Medicine Practice (PCIM) and Spine Practice) and BMC Springfield, MA in order to develop these models. We used various data analysis to understand the underlying system dynamics, which enabled us to identify the bottlenecks in the healthcare delivery system and to decide on which areas to focus on and improve.

(4) Another crucial and unique element of our projects is the close collaboration we have with stakeholders. As discussed in Retsef Levi's response to Green's editorial, to have an impact on healthcare delivery, the tools developed have to incorporate organizational specifics into the model and there should be institute-level

collaborations as well as the willingness of researchers to explore and understand the cultural environment (Levi and Prestipino [2012]). In our projects, we were able to establish this essential connection with a group of interdisciplinary collaborators.

1.3 Methodology

Using mathematical models to solve problems in clinical settings is a very complex process. The assumptions supporting the mathematical models need to be clinically realistic. These projects involved the interactive face to face process of reviewing and comparing mathematical assumptions and the clinical assumptions. This process is time consuming but essential to validate the models. The type of mathematical models we have used in formulating these clinical problems in the following chapters are:

1.3.1 Mixed integer programming

Optimization is a subject that deals with the problem of minimizing or maximizing a certain objective function in a finite dimensional Euclidean space, which is usually determined by functional inequalities. It involves achieving a single objective or multiple objectives by determining the values of the decision variables. Mixed integer programs are a subset of mathematical optimization models in which some or all of the decision variables are restricted to be integers. Since in most of the real world problems the decision variables are positive integers this area of optimization has been widely researched. Integer programming has been used for scheduling since World War 2 as George Dantzig mentions (Freund [1984]). IP models have been

extensively used in healthcare, especially in areas like hospital location problems, medical staff and patient scheduling problems (Cao and Lim [2011]). In Chapter 5 we model the spine surgery problem with an integer program.

1.3.2 Stochastic modeling

A stochastic model is a tool that makes use of probability distributions and theory to model real-world situations, by allowing for random variation in inputs over time, typically estimated from a historical data.

1.3.2.1 Queuing theory and modeling

Queuing theory was developed in 1904 by A. K. Erlang to determine the capacity requirements in a call center (Brockmeyer et al. [1948]). Queuing theory concerns the study of wait lines (Gross and Harris [1985]). It can translate customer arrival characteristics and service patterns into measures of waiting experienced by the customers like, average waiting time or the chance that customers will be delayed in the service process. Delays result from the mismatch between demand and service capacity. And as healthcare is riddled with delays, it is an ideal application area (Green [2011]).

Unlike simulation models, queuing models do not require a lot of data and have simple closed form expressions for many performance measures. Thus they are much faster to run and ideal for comparing different scenarios. However, the models developed incorporate many assumptions in order to develop closed-form expressions. Green describes the basic queuing principles in her chapter (Green [2011]).

In our paper, Liu, N. and Ozen, A. and Balasubramanian, H.J. [2014], we measure access to care by appointment delays (i.e., wait time) and operationalize continuity of care by the percentage of patients who see their own primary care providers (Chapter 3). Since we are interested in studying the relationship among panel size (which is directly related to patient appointment demand), provider service capacity and patient appointment delays, queuing theory is an ideal tool for this setting.

1.3.2.2 Newsvendor model

Newsvendor problems solve the optimal size of a single order to be placed before observing the stochastic demand when there are overage and underage costs. It is originated from the problem a newsvendor faces every day, when trying to decide how many newspapers to stock on a newsstand before observing the demand. The overage cost results from ordering too much, and underage costs from ordering too little of a perishable item.

Newsvendor models have been applied to capacity planning decisions for single period stochastic demand problems. It has been used in healthcare capacity decisions as well, for instance, Green et al. [2007a] use a newsvendor model approach in their paper to determine the relationship between the size of a physician's panel and the overflow frequency, where overflow frequency is the probability that the demand will exceed the available physician capacity. The demand for a panel of patients is a binomial random variable. Based on what the capacity of a physician is, the probability of overflow can then be easily calculated by using the cumulative distribution for the binomial random variable. The approach we use in Chapter 2 (Ozen and Balasubramanian [2013]) is closest to the modeling framework of Green et al. [2007a].

Their newsvendor like approach is extended to include case-mix and also to establish the interrelationship between multiple physicians working in a group practice.

1.3.2.3 Simulation (discrete event and object-oriented simulation)

Simulation is one of the most common methodology used in healthcare applications of operations research. One of the major reasons is that, computer simulation is a method that allows experimenting on a system while avoiding all the complications, like adding new staff, buying new expensive resources (Carr and Roberts [2011]). Both discrete event and object oriented simulation models have been widely used in healthcare since 1978 (Hancock and Walter [1979]). It has been applied to many areas, including the management of capacity like staff scheduling and admissions scheduling in operating rooms, outpatient clinics and ambulatory care (Forsberg et al. [2011]). The main difference between discrete event and object-oriented simulation is that discrete event simulations execute time-ordered events when a system changes state. On the other hand, object-oriented simulation requires the design and implementation of the objects, where objects are instances of classes, which are composed of properties and methods.

In our project with BMC (Chapter 4), we were able to both acquire data and establish active stakeholder participation, when developing the simulation model. This is essential, since in order to "sell" the results of the simulation to several parties the "stakeholders" need to be active participants in all phases of the simulation project (Carr and Roberts [2011]).

1.4 Dissertation Overview

This dissertation consists of six chapters. Chapters 2 and 3 of the dissertation is related to the primary care aspect of the capacity planning problem. In Chapter 2, I formulate an integer non-linear program for redesigning panels in a primary care group practice. In Chapter 3, our objective is to develop methods for evaluating access to care and continuity of care in commonly-used primary care delivery models adjusted for case-mixes; and to study how these two system performance measures change under panel (re)design and provider capacity pooling. Chapter 4 focuses on inpatient bed capacity planning with a goal of providing guidelines, particularly developing effective discharge policies, on how hospitals should manage their inpatient bed capacity in the presence of demand, discharge and length of stay (LOS) variability. Chapter 5 is on surgical care to create an optimal spine surgery master schedule by considering multiple objectives related to utilization, overtime, and financial performance. Lastly, the final chapter proposes future directions for the current research problems.

The dissertation is based on the following papers: Ozen and Balasubramanian [2013] and Liu, N. and Ozen, A. and Balasubramanian, H.J. [2014], Ozen, A. and Marmor, Y. and Rohleder, T. and Balasubramanian, H. and Huddleston, J. and Huddleston, P. [2014a], Ozen, A. and Marmor, Y. and Rohleder, T. and Balasubramanian, H. and Huddleston, J. and Huddleston, P. [2014b], Ozen, A. and Balasubramanian, H. and Roche, J. and Samra, P. and Ehresman, M. and Li, H. and Fairman, T. [2014] which are still under review.

CHAPTER 2

THE IMPACT OF CASE-MIX ON TIMELY ACCESS TO APPOINTMENTS IN A PRIMARY CARE GROUP PRACTICE

2.1 Introduction

Primary care providers (PCPs) are typically the first point of contact between patients and health systems. They include family physicians, general internists, and pediatricians. A primary care physician's (PCP) *panel* refers to the patients whose long term care she is responsible for. Over time, the PCP becomes familiar with the patients in her panel and is therefore able to deliver more informed and holistic care, with a focus on prevention. This long-term patient-physician relationship, also termed as *continuity of care* is one of the hallmarks of primary care.

The benefits of continuity for both patients and physicians have been well documented in the clinical literature. Gill and Mainous [2010] point to several studies

which show that patients who regularly see their own providers are 1) more satisfied with their care; 2) more likely to take medications correctly; 3) more likely to have their problems correctly identified by their physician; and 4) less likely to be hospitalized. Continuity and coordination are especially important for vulnerable patients with a complex medical history and mix of medications (Nutting et al. [2003]).

In practice continuity translates to maximizing patient-PCP matches when appointments are scheduled. But the ability of a PCP to provide continuity and timely access depends on 1) panel size, or the number of patients in her panel; and 2) case-mix, or the type of patients in the panel. For example, a panel consisting of mostly healthy patients will have a very different appointment burden compared to a panel consisting mostly of patients with chronic conditions.

In this paper, we characterize the interrelationship between panel size, case-mix and the individual capacities of physicians working in a group practice. This is done by measuring the *overflow frequency* of the physicians in relation to each other. The overflow frequency is the probability that the demand from a physician panel (i.e. patient requests for appointments in a day) will exceed the physician's capacity (i.e. the number of appointment slots a physician has available in a day). A high overflow frequency for a physician implies that patients in the panel will be unable to access their physician in a timely manner and are as a result more likely to visit an unfamiliar physician or emergency room. Thus a high overflow frequency implies that both timely access and continuity of care are compromised.

The consideration of panel size and case-mix in this paper, is particularly relevant given the acute shortage of PCPs in the United States. The demand for primary care

continues to grow as the population ages and the prevalence of chronic conditions increases. Our approach allows practices to quantify their current supply and demand imbalances and use available capacity in the most efficient manner possible. Case-mix is an important consideration given that patient demographics and care needs vary from community to community and from one geographic region to another.

The analysis presented in this paper is at the aggregate planning level, where a practice has to decide how many and what type of patients are appropriate in each panel to ensure patients have adequate levels of access and continuity. In the long term, if imbalances in workload exist among the physicians, a practice may be interested in *redesigning* panels – that is in changing the size and case-mix of individual physician panels so that each physician's capacity is in balance with her demand. While this involves changing existing panel configurations, opportunities for redesign arise constantly in primary care (more details in Section 2.5). For example, new patients may join the practice, existing patients may move from the area, and patient preferences about their PCP may change over time. On the capacity side, a physician may leave the practice or retire, with the result that patients in that physician's panel now need to be reassigned. In residency practices found in academic medical centers, the turnover of residents every year provides constant opportunities for panel redesign. We discuss the feasibility of panel redesign in greater detail in Section 2.5.

We propose an integer non-linear programming formulation for redesigning panels in a group practice. The goal is to minimize the maximum overflow frequency over all physicians. Rather than prescribe exactly what practices should do, we derive

analytical results to benchmark a practice's current performance. Then the analytical results are used to motivate heuristics, which will allow practice managers to: 1) test various redesign options and, 2)infer which options are the least disruptive. An important advantage of our approach is that all our analytical results can be implemented in Excel and used for aggregate level planning and panel management decisions.

The rest of the paper is organized as follows. In Section 2.2, the relevant literature is reviewed and in Section 2.3, we explain the modeling of case-mix. We motivate the panel redesign problem using an example involving 4 physicians in Section 2.4. The feasibility of panel redesign in practice is discussed in Section 2.5. Section 2.6 contains all the mathematical details and analytical results related to the panel redesign formulation. In Section 2.7, the heuristics are described. In Section 2.8, we explain how we used patient and panel data from the Primary Care Internal Medicine (PCIM) practice in Rochester, Minnesota to create four test practices to demonstrate the results. Section 2.9 summarizes the conclusions and explains the implications of our results for practices.

2.2 Literature Review

Appointment scheduling in healthcare is an active and growing area of research. Over the last decade, the advanced access paradigm, made popular in clinical journals by Mark Murray (Murray and Tantau [2000]; Murray and Berwick [2003]; Murray et al. [2007]), attempted to promote same-day access for patients. In traditional appointment systems, appointments are allowed to be booked into the future, whereas in

advanced access this is discouraged. All appointments, regardless of their nature and urgency of request, are to be seen the same day by the patient's PCP. In practice, most clinics follow a blend of traditional and advanced access scheduling. Clinical necessities (follow-ups for chronic conditions) and patient preferences require practices to allow the future booking of appointments, while at the same time enable same-day access for acute needs. Yet, whatever appointment system or blend a practice may follow, effective access is possible only if the panel sizes of the physicians and their case-mixes are in balance with the available capacity, and the impact of variability is adequately addressed.

The operations research literature has in the last decade tackled a number of aspects related to appointment scheduling using stochastic optimization approaches. This includes an analytical comparison of traditional and advanced access appointment systems (Robinson and Chen [2010]); the impact of no-shows (LaGanga and Lawrence [2007], Muthuraman and Lawley [2008], and Chakraborty et al. [2010], Liu et al. [2010]); the importance of considering patient preferences (Gupta and Wang [2008], Wang and Gupta [2011]); and capacity allocation methods that allow practices to offer a blend of prescheduled (non-urgent) and same-day (urgent) appointments (Balasubramanian et al. [2011] and Qu et al. [2006]).

We reiterate that the analysis presented in this paper is at the aggregate level. Thus we only focus this review on the papers most relevant to our work on panel size and case-mix. Murray and Berwick [2003] proposed six steps for clinics to implement advanced access. An important message of this work is that the primary lever for demand is the number of patients in a physician's panel. Murray et al.

[2007] provide a simple algorithm to calculate the "right" panel size for physicians. Murray et al. [2007] also mention other factors that might affect the workload of physicians like gender and age (panel case-mix) but do not provide any quantitative analysis. While the paper provides clinics with easily implementable policies to realize advanced access by resizing panels, there is no discussion on the impact of variability, an important factor in appointment scheduling.

Green and Savin [2008] use queuing models and simulation to demonstrate the impact of panel size on the no-show rate, physician utilization, and the probability of getting a same-day appointment. They find that the backlog of appointments grows with panel size and as a result the no-show rate does as well, since patients booked well into the future will have a greater probability of no-show.

In Green et al. [2007a], a newsvendor like model is proposed to determine the relationship between the size of a physician's panel and the overflow frequency. Overflow frequency, as stated in Section 2.1, is the probability that the demand will exceed the available physician capacity. They assume that each patient in the panel has a probability p of requesting an appointment on any given day. This probability can be estimated from historical visit rates. Since each patient requests independently of each other, the demand for a panel of patients is a binomial random variable. Based on what the capacity of a physician is, the probability of overflow can then be easily calculated using the cumulative distribution function of the binomial distribution.

The approach we take is closest to the modeling framework of Green et al. [2007a]. Their newsvendor like approach is extended to include case-mix and also to establish the interrelationship between multiple physicians working in a group practice. We

first extend the binomial framework for modeling demand to consider different classes of patients. In our model, case-mix is represented by the number of simultaneous chronic conditions a patient has (more details in Section 2.3). Next, the overflow frequency is used as a measure of access, and then theoretical results are developed that will allow a group practice to benchmark their current performance. Finally we develop simple heuristics that will allow practices to test long-term panel redesign scenarios. The results are demonstrated using panel data from the primary care internal medicine (PCIM) practice at Mayo Clinic.

2.3 Patient Classification

Patients can be characterized by various attributes, such as age and gender and the chronic conditions afflicting the patient. Our interest is in attributes that play an important role in determining the distribution of visits. In addition to operational and capacity planning reasons, patient classification can be useful for clinics because they enhance a practice's understanding of its population and disease trends, and allow it to design its care models effectively. Barbara Starfield's seminal work about ACGs (Ambulatory Care Groups) argued that understanding the role of patients' clinical complexity in care utilization forms the cornerstone for effective resource planning and determining payment methods in healthcare (Starfield et al. [1991]).

What classifications are the most effective in predicting appointment request rates? Age and gender is the simplest patient classification in absence of other data, yet is generally effective (Murray et al. [2007], Balasubramanian et al. [2010]). In this paper, the number of simultaneous chronic conditions a patient has is used

as a predictor of the number of visits. In clinical parlance, these conditions are *comorbidities*. Our choice is based on the following reasons. First, comorbidity counts have clinical relevance and are widely accepted by the primary care practices we have interacted with. Focusing on all comorbidities of a patient is more holistic than focusing in isolation on specific chronic conditions, and primary care was conceived to be a holistic approach rather than a disease specific approach. Secondly, our categorization has been used both in literature and practice. Naessens et al. [2011] show that the number of simultaneous chronic conditions is a strong predictor of the number of office visits. Comorbidity counts have also been used in the new payment scheme for primary care proposed by the Minnesota Department of Health (MDH [2010]). Finally, statistical analysis of the patient level data from Mayo Clinic (using classification and regression trees, CART) revealed the count of comorbidities as the strongest predictor of appointment request rates.

We note, however, that the models proposed in this paper can be applied to any patient classification. While patient classification is important, the central theme of this paper is not to find the "best" classification. Rather, it is to show the impact of patient classes on access measures. To illustrate the impact of comorbidity counts, we analyzed the patient population (around 27,000 patients) empanelled at the Primary Care Internal Medicine Practice (PCIM) at the Mayo Clinic in Rochester, Minnesota. Examples of commonly observed chronic conditions in patients included hypertension, depression, diabetes, osteoporosis, urinary tract infections, hyperlipidemia, coronary artery disease and otitis. We divided patients based on the number

of comorbidities they had. In all there were 8 patient categories as patients with more than 7 comorbidities was extremely rare.

Table 2.1 below summarizes the number of patients, average number of visits and standard deviation for each comorbidity count category, based on historical visits in PCIM. Clearly, not only does the mean number of visits increase with the number of comorbidities, the standard deviation does as well. The standard deviations are higher than the means, suggesting significant variation in visit rates within each comorbidity count category.

Table 2.1: Mean and standard deviation of visits in 2006, for patients with different counts of comorbidities

# of Comorbidities	**# of patients**	**avg visits/pat/year**	**Std Dev.**
0	6524	1.72	2.88
1	6980	2.74	4.56
2	5819	3.82	6.25
3	4179	5.16	8.56
4	2370	6.82	9.95
5	989	7.67	10.72
6	346	9.62	13.14
7	84	11.17	13.39

2.4 Example of 4 Physicians

In this section, we demonstrate the impact of case-mix using a simple simulation. In the general case, there are $j = 1, ..., J$ physicians in the practice. Suppose all patients empanelled in a practice have been categorized into $i = 1, ..., M$ patient classes. A patient of category i has a probability p_i of requesting an appointment on a given day. This probability will be higher for patients with multiple chronic conditions than for relatively healthy patients (see Section 2.8 for the exact values

and how these probabilities are calculated). Next, suppose n_{ij} denotes the number of class i patients in physician j's panel. The total demand for the physician is the sum of the demand from each patient class. The demand from each patient class is a binomial random variable – with n_{ij} patients in patient class i and probability of class i patient requesting on a given day being p_i.

The p_i and n_{ij} values are used to generate binomial data realizations using random sampling and thereby simulate the total demand for each physician. If we know the total daily appointment slots a physician has available in a day, then the simulation can be used to calculate the utilization, overflow frequency, and the expected overflow for each physician. Utilization is simply the expected total demand divided by the total daily slots a physician has available in a day. Overflow frequency is the fraction of total realizations (each realization can be thought of as a day) in which the patients' visit requests for the day exceed the available capacity of the physician. Expected overflow is the average patient backlog (unfulfilled demand) at the end of each day.

As an example, consider the results of the simulation for four PCIM physicians at Mayo Clinic. The physicians have approximately the same panel size (around 1060 patients), but different case-mixes: different patient numbers in the 8 comorbidity count categories. The panel compositions of each physician are shown in Table 2.2, as are the overflow frequency, expected overflow and utilization. All four physicians have a capacity of 17 slots. We use 10,000 realizations.

Notice that Physician 3 and Physician 1 have relatively high utilizations, overflow frequencies and expected overflows. This is because they have more patients with

two or more comorbidities in their panels, and these patient groups generate a higher number of visits. High overflows result in 1) patients seeing an unfamiliar physician or visiting an emergency room (loss of continuity), or 2) longer wait times to secure an appointment (lack of timely access).

Table 2.2: Case-mix, panel size and performance measures for 4 physicians each with a capacity of 17 appointment slots per day, where OF: overflow frequency; EO: expected overflow; Util: Utilization

Physicians	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**Panel Size**	**OF**	**EO**	**Util**
P1	260	249	226	161	108	42	14	3	1063	30%	3.64	92%
P2	299	293	212	147	77	26	6	1	1062	22%	0.94	87%
P3	214	253	223	177	115	44	21	5	1053	35%	7.36	95%
P4	290	296	218	145	84	27	12	5	1077	18%	1.48	83%

These results suggest that, in addition to using panel size, clinics may benefit by making capacity and allocation decisions based on case-mix. In the face of high overflows, physicians generally work longer hours. But this is not an appealing option, especially in primary care where reimbursements are low, and where more and more physicians are experiencing emotional exhaustion because of the number of patients they have to see (Bodenheimer and Pham [2010]). The long-term option for practices is to *redesign panels*. This means changing case-mix proportions by reassigning patients across panels so that each physician's demand is in balance with her capacity.

2.5 Feasibility of Panel Redesign

Before describing the panel redesign formulation, it is important to discuss how feasible or useful such a framework is to practices, individual physicians and patients.

Redesigning panels implies changing existing patient-physician relationships, and there appears to be a paradox. To improve timely access and continuity in the long run a practice has to invest in the short term disruption of existing patient relationships. It is natural therefore to ask: how realistic is redesign in practice?

The feasibility of redesign would be a very valid concern if each patient in the panel was very loyal to the physician and had spent many years visiting the physician. Enforcing a break in that relationship would not be satisfactory to either the patient or the physician. But in practice, a panel is a lot more fluid. While there exist many patients who have spent years with the physician (we do not recommend that these relationships be disrupted), there also exist patients who are newly registered or are as yet uncommitted to their physician even though they have been assigned to a panel. It is these patients who would be amenable to redesign.

For example, in order to improve access to care, continuity and care coordination, Group Health Practice of Seattle recently reduced panel sizes from 2300 per physician to 1800 per physician (Reid, R. J. and Fishman, P.A. and Yu, O. and Ross, T. R. and Tufano, J. T. and Soman, M.P. and Larson, E.B. [2009]). They hired new physicians and reassigned 500 patients per physician to either new physician or physicians who had available capacity. Patients were invited to an open house to meet their new physicians and surveys were used to identify patients who were willing to change their PCP.

In their papers, Reid, R. J. and Coleman, K. and Johnson, E. A. and Fishman, P.A. and Hsu, C. and Soman, M.P. and Trescott, C. E. and Erikson, M. and Larson, E.B. [2010] and Coleman et al. [2010] analyze the Group Health Clinic after the im-

plementation. They used survey-based measures to quantify patient satisfaction and staff burnout. The results of the implementation were: 1) Staff burnout decreases since they find that emotional exhaustion becomes less frequent for physicians; 2) Patients' experience improves in terms of access to care and doctor-patient interactions (and this manifests itself in 29% fewer ED visits and 11% fewer hospitalizations); 3) During the reassignment, when physicians are given the chance to choose patients to keep in their panel, they prefer the elderly and sicker patients, who create a greater density of visits and need more continuity; and 4) Reassigned patients use primary care less, but there is no significant increase in their use of the ED.

While Group Health seems to have successfully achieved its redesign to improve patient centeredness, access and continuity, their reassignment of patients does not seem to have followed a quantitative basis. For example, how did the practice decide that 500 patients per physician (more than 20% of the original panel size of 2300) had to be reassigned? Could fewer patients have been reassigned or do panel sizes need to be even smaller? Quantitatively capturing the beneficial effects of redesign and the impact on the number of patients affected – which is the focus of this paper – will help individual physicians and the practice as a whole to make the choices that are most appropriate for them.

Indeed our experimental results (see Section 2.8.2) based on the primary care internal medicine practice (PCIM) at Mayo Clinic suggest that panel redesign will affect at most $5 - 8\%$ of the total patients (250 patients out of 4300 total) in the practice. Furthermore, the number of patients affected can be as low as 2% (less than 100 out of 4300 total). So the very large majority of patient physician relationships

will remain unaffected. Yet, the improvements in overflow frequency due to redesign are significant for the overburdened physicians in the PCIM practice. There is thus a strong incentive for overburdened physicians to consider redesign, since it improves access measures for their patients.

Furthermore, as Balasubramanian et al. [2010] argue, redesign does not need to be carried out instantly as in the Group Health case, but can be achieved by most practices in the long term. Every practice has a natural attrition rate as well as a group of new patients wanting to join the practice. Patients' comorbidities can change over time as well. Retiring physicians will need to transition their patients to newly hired physicians. These rates could be used, over time (a period of 1-2 years or perhaps more) to adjust case-mixes so that timely access and continuity are improved. Indeed we view the framework of this paper not as a strict prescription that dictates what practices should do. Rather it is an assessment tool, which practices can use to benchmark their current access and continuity levels on a quarterly or yearly basis and use whatever leverage they have to change panels.

2.6 The Panel Redesign Formulation (PRF) and Analytical Results

In this section, a mathematical formulation is provided to redesign physician panels in a multi-physician practice to minimize the maximum overflow frequency. We choose overflow frequency since it is a more tractable non-linear objective function than the expected overflow. It also allows us to derive properties that eventually allow near optimal solutions to be reached using simple heuristics. Later in the results section,

the positive correlation we have already observed between overflow frequency and the expected overflow, will be seen again.

We choose a minimax objective function over a summation function because even if the sum of overflow frequencies over all physicians in the practice is minimum, some physicians may still have higher overflow frequencies in relation to others. This will eventually lead to redirections to unfamiliar physicians and hence a loss of continuity. The minimax function, on the other hand, will ensure to the extent possible that each physician's panel demand is in balance with her capacity. We will also see in this section that identical overflow frequencies for all physicians does not mean that physician panels have to be identical in their case-mix proportions.

As discussed in the Section 2.4, n_{ij} denotes the number of patients from patient class i in physician j's current panel. The n_{ij} values over all J physicians and all M patient classes together describe the current panel design. However, the practice would like to *redesign* panels, that is determine new allocations from each patient class i to each physician panel j to minimize the maximum overflow frequency. Let x_{ij} be the number of patients to be assigned from patient class i to physician j. The constraints are that x_{ij} values should be integer and that all patients from each class have to be allocated, $\sum_{j=1}^{J} x_{ij} = N_i, \forall i = 1, ..., M$. Here N_i is the total number of class i patients (or category i patients) in the practice.

As before, the probability that a patient of class i requests for an appointment on any given day is p_i. If we assume that patients request independently of each other then the total demand for physician panel j from patient class i after reassignment is a binomial random variable with mean $x_{ij}p_i$ and variance $x_{ij}p_i(1 - p_i)$. If we

take the sum over all M patient classes, the mean and standard deviation of the total demand arising from physician j's panel are given by: $\mu_j = \sum_{i=1}^{M} p_i x_{ij}$ and $\sigma_j = \sqrt{\sum_{i=1}^{M} p_i(1-p_i)x_{ij}}$, respectively, $0 < p_i < 1$. Note that both the mean and standard deviation depend on the case-mix distribution given by the x_{ij} values for the physician. If panel sizes are sufficiently large ($>$ 800-1000 patients), the total demand is the sum of as many Bernoulli random variables, and is likely to be well approximated by a normal distribution. We verified this statistically by applying the Kolmogorov-Smirnov (KS) goodness of fit test. The test was applied to total demand data generated using 10,000 random samples from the binomial demand distributions corresponding to the individual patient categories.

Let C_j denote the capacity of the physician, the total daily slots that she has available in a day. Then Z_j, the standard normal Z-score for physician j, is given by: $Z_j = \frac{C_j - \mu_j}{\sigma_j}$. Intuitively, the Z-score gives the number of standard deviations that the capacity is distant from the mean of the panel demand. If the percentile of the standard normal distribution is denoted by Φ, then the probability of overflow for physician j, O_j, is $O_j = 1 - \Phi(Z_j)$. The greater the positive distance between C_j and μ_j and the smaller the σ_j, the greater the Z_j value and the lower the overflow frequency O_j.

The goal is to optimize x_{ij} allocations to minimize $\max\{O_1, O_2, ..., O_J\}$ – that is minimize the maximum overflow frequency over all physicians in the practice. The formulation is summarized below. We call it the panel redesign formulation (PRF).

$$(PRF) \qquad \min_{x_{ij}}\{\max\{O_1, O_2, ..., O_J\}\} \tag{2.1}$$

$$s.t. \qquad O_j = 1 - \Phi\{\frac{C_j - \mu_j}{\sigma_j}\}, \forall j = 1, ..., J \tag{2.2}$$

$$\mu_j = \sum_{i=1}^{M} p_i x_{ij} \quad \forall j = 1, ..., J \tag{2.3}$$

$$\sigma_j = \sqrt{\sum_{i=1}^{M} p_i(1 - p_i) x_{ij}} \quad \forall j = 1, ..., J \tag{2.4}$$

$$\sum_{j=1}^{J} x_{ij} = N_i \quad \forall i = 1, ..., M \tag{2.5}$$

$$x_{ij} \geq 0 \text{ and integer } \forall (i, j) \tag{2.6}$$

Note that PRF is an integer non-linear program. The formulation is described visually in Figure 2.1. The total mean and variance of the entire patient population given by $\mu_{total} = \sum_{i=1}^{M} p_i N_i$ and $\sigma^2_{total} = \sum_{i=1}^{M} p_i(1-p_i)N_i$. The allocation problem is all about optimally partitioning the total population mean, μ_{total}, and variance, σ^2_{total} to individual physicians in the practice. The lever through which the partitioning is achieved are the x_{ij} values. The means and variances are not allocated independently of each other but are tied to the x_{ij} allocations. In other words, O_j, μ_j and σ_j will all increase (decrease) together when x_{ij} increases (decreases) for any $i = 1...M$.

Clearly, the maximum overflow will always be minimized if all the Z-scores and corresponding overflows can be made equal. Even if they cannot be made exactly equal, the differences in the overflows will be small enough to be negligible for large panel sizes. In other words, there is sufficient granularity in large panels (> 800 patients) to smooth the overflows in the practice.

Consider, first, the equal capacity case, $C_1 = C_2 = C_3.... = C_J$, which is relatively easy to understand. Since the physicians are all identical, then any allocation in

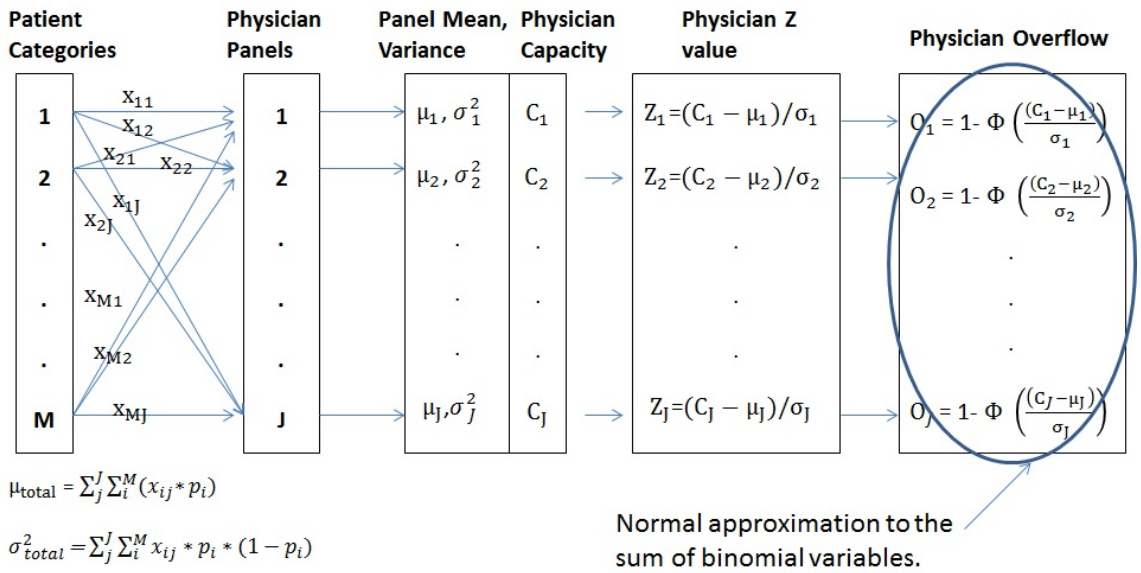


Figure 2.1: A visual summary of the panel redesign problem to minimize the maximum overflow

which $\mu_1 = \mu_2 = ... = \mu_J = \mu_{total}/J$ and $\sigma_1^2 = \sigma_2^2 = ... = \sigma_J^2 = \sigma_{total}^2/J$ will minimize the maximum overflow frequency. Note the above statement refers to a *set of allocations*, not a particular one – the optimal overflow can be reached in multiple ways.

A special case is the allocation where each physician j gets the same number of patients from each category i. Mathematically, $x_{ij} = N_i/J, \forall i, j$. To maintain integrality of the decision variables in such a symmetric allocation, the number of patients in each category i should be a multiple of the number of physicians J in the practice. Even if this condition does not hold true, the general idea is that all physicians have nearly identical panel compositions. A symmetric allocation has practical benefits. PCPs are the generalists of healthcare. Their training allows them

to treat a wide variety of patients, ailments and chronic conditions. The $x_{ij} = N_i/J$ allocation maximizes diversity of patients in physician panels. This is especially important for panels of primary care residents in academic medical centers. Patient and diagnostic diversity is an essential education and training objective of a resident. Similarly private practices with a large number of relatively new physicians might benefit from introducing diversity in panels.

Practices, however, do not have to follow such symmetric allocations. Panels tend to grow more organically over time. In the interest of not disturbing existing patient-physician relationships, a practice may choose other allocations that are asymmetric yet in a manner that the overflows turn out to be identical. Thus, although the structure of allocations in the equal capacity case is obvious, the subtle point is that there are multiple optimal solutions. We revisit this theme again in the heuristics and results section. One of our objectives there is try to redesign panels with the minimum possible disruption to existing panels.

We next consider the more general unequal capacity case: $C_1 \neq C_2 \neq C_3 .... \neq C_J$. In academic medical centers, where physicians have research responsibilities, the unequal capacity case is more prevalent. But even in non academic small practices, with 3 or 4 physicians on staff (where majority of primary care in the U.S. is delivered), physicians will often have different schedules or may work only part time. Physicians on the path to retirement also may gradually reduce their work hours.

2.6.1 The unequal capacity case

When the physicians have different number of slots available every day, it would seem appropriate to allocate patients keeping in mind the capacity a physician has. Greater capacity would imply a greater share of μ_{total} and σ^2_{total}. However, the difficulty is in determining precisely *how much greater* that share should be for an optimal allocation. Let C be the total capacity of the clinic – total slots the clinic has available on a typical workday. Therefore $C = C_1 + C_2 + ... + C_J$. An allocation in proportion to the capacity is given by: $x_{ij} = (C_j/C) * N_i$ for all i and j. In other words, the number of patients from each category is proportioned in the ratio of an individual physician's capacity to the total clinic capacity. This seems an intuitive way of allocating patients and is an extension of the equal capacity case where each physician was assigned the same number of patients.

However, the allocation $x_{ij} = (C_j/C) * N_i$, while likely to be a good heuristic, is not guaranteed to give the optimal solution (specific examples in Section 2.8). This is because while the allocation of patients from each patient class increases linearly as the capacity increases, the objective function changes non-linearly. Indeed, a simple closed form expression for the optimal allocation, as described in the equal capacity case, may not be possible. It may be possible to solve PRF (at least numerically) by relaxing the integrality constraints on x_{ij}. However, rather than choosing this course, we approximate the optimal objective. This will give practices a reference or a target overflow frequency, O_{ref} to aim for when they redesign panels. We show that for all practical purposes O_{ref} is a good surrogate for the optimal overflow frequency O_{opt}. A practice can use O_{ref} to test various redesign options (multiple ways of reaching

the optimal value), and choose whatever works best for them. This approach is less prescriptive than solving the non-linear program exactly to determine x_{ij} values. Furthermore, the calculation of O_{ref} can be achieved using an Excel spreadsheet, and therefore will be easy to implement in practice.

2.6.2 Deriving the reference overflow O_{ref}

Our method relies on relating overflow of individual physicians in the optimal allocation to the overflow of a hypothetical "combined physician". This combined physician (CP) is simply the aggregated system. In other words, the combined physician has a capacity of $C = C_1 + C_2 + ... + C_J$, a mean demand equal to μ_{total} and variance equal to σ^2_{total}. In such a practice, a physician can see the patients of any other physician – there is thus no concept of continuity. The standard normal value corresponding to the combined physician, Z_{CP} is given by:

$$Z_{CP} = \frac{C - \mu_{total}}{\sqrt{\sigma^2_{total}}} \tag{2.7}$$

Notice that the above expression can be easily obtained independently, without any knowledge of the x_{ij} values in the optimal allocation. We shall next try to relate the Z_{CP} value to the standard normal value Z_j for each physician j in an optimal allocation. Suppose μ_j, σ_j and Z_j represent the mean, standard deviation and Z value for physician j in an optimal allocation. For sufficiently large panel sizes, we know that the overflows of the physicians in an optimal allocation are approximately equal, which implies that the Z_j values will be approximately equal as well. So it is reasonable to write $Z_{opt} = Z_1 = Z_2 = Z_3 = ... = Z_J$. More precisely:

$$Z_{opt} = Z_j = \frac{C_j - \mu_j}{\sigma_j}, \forall j \tag{2.8}$$

$$\sigma_j * Z_{opt} = C_j - \mu_j, \forall j \tag{2.9}$$

If we add all the J equations, one for each physician, based on the equality above, we get:

$$\sum_{j=1}^{J} \sigma_j * Z_{opt} = \sum_{j=1}^{J} C_j - \sum_{j=1}^{J} \mu_j$$

$$Z_{opt} = \frac{\sum_{j=1}^{J} C_j - \sum_{j=1}^{J} \mu_j}{\sum_{j=1}^{J} \sigma_j} = \frac{C - \mu_{total}}{\sum_{j=1}^{J} \sigma_j} \tag{2.10}$$

From the expression for Z_{opt} and Z_{CP} (see equations 7 and 8) we have the following result.

$$Z_{opt} = \frac{Z_{CP}}{R}, \text{where} \quad R = \frac{\sum_{j=1}^{J} \sigma_j}{\sqrt{\sigma_{total}^2}} \tag{2.11}$$

Note that since $\sigma_{total}^2 = \sum_{j=1}^{J} \sigma_j^2$, we can rewrite R as:

$$R = \frac{\sum_{j=1}^{J} \sigma_j}{\sqrt{\sum_{j=1}^{J} \sigma_j^2}} \tag{2.12}$$

Notice that $R \geq 1$. This is because the sum of J positive numbers (the numerator of R) is always greater than the square root of the sum of squares of the J numbers (denominator of R). This means that $Z_{CP} \geq Z_{opt}$. The equality is tight when $R = 1$ (i.e., the extreme case where one physician has all capacity and all demand, while all

the others have none). We can also derive an upper bound on R. The upper-bound $R = \sqrt{J}$ is realized when all the J numbers involved in the expression are equal, that is $\sigma_1 = \sigma_2 = ... = \sigma_J$. We define $Z_{ref} = \frac{Z_{CP}}{\sqrt{J}}$. If the capacities of the physicians are equal, then $Z_{opt} = Z_{ref}$ and if the capacities of the physicians are unequal, we have $\frac{Z_{CP}}{R} \geq \frac{Z_{CP}}{\sqrt{J}}$, which implies $Z_{opt} \geq Z_{ref}$.

Intuitively, R captures the decline in variability when demands and capacities are aggregated (the well known aggregation effect). The decline is highest when each physician has the same variance (and standard deviation). As physician panels become more and more unequal with regard to the variances allocated to them, R starts to approach 1 and Z_{CP} starts to approach Z_{opt}. Indeed, to calculate the optimal Z_{opt}, we do not need to know the exact standard deviation values of the individual physicians. But we need to know how the standard deviations of the J physicians stand in relation to each other – that relationship is captured by R. From the above analysis, the following key result is derived:

$$Z_{CP} \geq Z_{opt} \geq \frac{Z_{CP}}{\sqrt{J}} \tag{2.13}$$

The overflows corresponding to the Z-scores above are given by $O_{CP} = 1 - \Phi(Z_{CP})$, $O_{opt} = 1 - \Phi(Z_{opt})$ and $O_{ref} = 1 - \Phi(Z_{ref})$ respectively. The relationship between the overflows can be described as follows:

$$O_{CP} \leq O_{opt} \leq O_{ref} \tag{2.14}$$

O_{CP} can be interpreted as the overflow of a practice that has no concept of panels. Any physician in the practice can see any of the total patients in the practice. There is no continuity. Such sharing however has the benefit of capacity pooling and hence O_{CP} is the best overflow a practice can achieve – it is the lower bound. O_{opt} on the other hand is the overflow of each physician assuming that the physicians do not share their patients at all. This provides perfect continuity but the benefit of capacity pooling is lost. Practices usually lie between these two extremes. Thus the difference between O_{opt} and O_{CP} measures *the price of continuity.*

While there is no exact method of computing O_{opt}, $O_{ref} = 1 - \Phi(Z_{ref})$ is used as a surrogate for the optimal overflow. It will be demonstrated that $O_{ref} - O_{opt}$ is fairly small for most cases found in practice. Indeed, for the equal capacity case $R = \sqrt{J}, Z_{ref} = Z_{opt}$ and therefore $O_{ref} = O_{opt}$: the reference value is exactly equal to the optimal value.

2.6.3 $O_{ref} - O_{opt}$ for common cases in practice

To characterize $O_{ref} - O_{opt}$ we must consider what values of R are reasonable in practice. Consider a 2-physician practice. When the physicians have identical capacities, we expect to see $\sigma_1 = \sigma_2$ in the optimal allocation and therefore $R = \sqrt{2} = 1.414$. The more unequal the physicians are with regard to their capacities, the more R starts to approach 1.

When the capacities of the two physicians are not equal, the optimal allocation is unknown. But the asymmetry in physician capacities can give us a hint of what the R value might be. Suppose one physician works full time and has 24 slots in a

day (assuming an 8 hour day with 3 patients per hour, a typical workload for PCPs), while the other physician works only 6 slots in a day. This asymmetry in capacities is perhaps the limit of what might be observed in a practice – seeing 6 patients a day (about 2-3 hours of work per day) is generally not common except in residency practices.

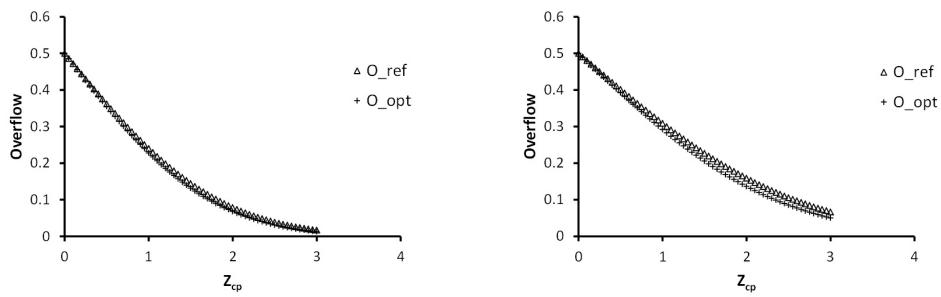
Although the optimal allocation of patients for the above case is not known, we can still state that the mean and variance allocated to the full time physician should be *roughly* four times that allocated to the quarter-time physician. This can be stated because, it is known that the mean and variance are tightly coupled through the x_{ij} values – they both increase and decrease together. So we have: $\mu_1 = 4\mu_2$ and $\sigma_1^2 = 4\sigma_2^2$. This gives us an R value of 1.34. So $R = 1.34$ represents (approximately) a fourfold variation in capacities for a 2-physician practice. R values smaller than this imply that one physician works a negligible amount of time daily in relation to the other. Capacities of 12 and 24 or 10 and 20 are more reasonable since since some physicians may work full time while others may work only for half a day. For such cases $R \geq 1.34$. In general, all practical 2 physician cases are well represented by $1.34 \leq R \leq 1.414$.

So a 2-physician practice which has $R = 1.34$ allows us to test the strength of our reference value O_{ref}. If O_{ref} approximates O_{opt} well for for this case, it will be even better for $R > 1.34$, which are more commonly observed.

As an example, suppose we find that $Z_{CP} = 1.0$ for a 2 physician practice with $R = 1.34$ (recall that Z_{CP} can be computed independently). If we do not know anything about the optimal allocation, our only option is to use the reference value,

$Z_{ref} = \frac{Z_{CP}}{\sqrt{J}} = \frac{1.0}{1.414} = 0.707$. The optimal value, using $R = 1.34$ is $Z_{opt} = \frac{Z_{CP}}{R} = \frac{1.0}{1.34} = 0.746$. It follows that the reference overflow and optimal overflow are $O_{ref} = 1 - \Phi(Z_{ref}) = 1 - 0.772 = 0.239$ and $O_{opt} = 1 - \Phi(Z_{opt}) = 1 - 0.745 = 0.2227$ respectively. The difference is within 1%.

Figure 2.2a below shows O_{ref} and O_{opt} as a function of Z_{CP}, which is varied from 0 to 3. The two lines are almost indistinguishable. At $Z_{CP} = 0$, when the aggregated demand equals the aggregated supply and the utilization is 100%, both O_{ref} and O_{opt} are 0.5. The prediction is exact. As the overflow decreases, O_{ref} and O_{opt} differ from each other, with O_{ref} always being larger, but the difference never exceeds 1.3 %.



(a) 2-physician case (b) 4-physician case

Figure 2.2: Comparison of O_{ref} and O_{opt} as a function of Z_{CP} for the 2-physician and 4-physician example

To further reinforce the point a 4 physician example is considered. Here we assume a sixteen-fold difference in capacities $C_1 = 4C_2 = 9C_3 = 16C_4$, which is an extreme limit on the capacity variation a practice is likely to have. Here the variance relationship will approximately be: $\sigma_1^2 = 4\sigma_2^2 = 9\sigma_3^2 = 16\sigma_4^2$. The R value for this setting is 1.825. We use $O_{ref} = 1 - \Phi(Z_{ref}) = 1 - \Phi(\frac{Z_{CP}}{sqrt(4)})$ as the reference value. If Z_{ref} works for well for this case, it will work even better for $1.825 < R \leq 2$. Figure

2.2b shows O_{ref} and O_{opt} as a function of Z_{CP} for the 4 physician example where $R = 1.825$. Here the difference between the two is slightly larger but O_{ref} is still within 2.5% of O_{opt}. We have thus shown that O_{ref} is good surrogate for the optimal overflow O_{opt} for practical cases.

2.6.4 Summary of contributions

In summary, the PRF formulation allows a practice to:

1) Benchmark the access performance of each physician in the practice with other physicians as well as the reference overflow value.

2) Capture the price of continuity (in terms of lost access). Specifically, the price of continuity for a practice is the difference between the reference or target overflow and the overflow of a practice in which all physicians together serve all the patients in the practice (no concept of a panel, but pooled capacity to meet the demand).

3) Quantitatively evaluate and arrive at the least disruptive way of redesigning panels, since achieving the reference overflow is possible in many different ways (multiple optimal solutions). This allows a practice to quantify the minimum number of patients whose current PCP assignments will be affected if redesign were to be implemented.

Our heuristics and results, described in the next sections, quantitatively demonstrate each of these contributions and provides the foundation for a spreadsheet-based decision tool for aggregate level panel management decisions in a group practice.

2.7 Heuristics

In the last section, we have seen how a reference or target overflow can be determined for a group of physicians, and that this value is a good proxy for the optimal overflow for most practical scenarios. In this section, heuristics are described that practices can use to switch patients between panels so that this target overflow is achieved. Since switching patients disrupts existing patient-PCP relationships, a practice will be keen to 1) minimize the number of patients that are switched; 2) ensure that patients with the greatest continuity needs (for example a patient with multiple chronic conditions) are not switched. As it is demonstrated with our heuristics, these two goals can be conflicting.

Before explaining our heuristics, it is important to note that we assume that patient categories are ranked in non-decreasing order, based on their p_i values, which determines the visit rate of that patient category. In our classification method for instance, zero comorbidity patients have the lowest visit rate, one comorbidity patients have the next lowest visit rate and so on.

To use the patient switching heuristics, practices start with an initial solution, for example the practice's current case-mix or current panel design. Next, the overflow value for each of the physicians is computed based on the initial solution. The physicians are ranked in decreasing order of their overflow values. A patient of the lowest visit category (the group with 0 comorbidities in our case) is then selected from the panel of the physician with the highest overflow and is now assigned to the panel of the physician with the lowest overflow. The overflow values for the two physicians are updated. If maximum overflow for the practice is greater than

the reference overflow value (calculated as described in the previous section), another patient from the lowest visit category is transferred. If the physician with the highest overflow has no more patients in the lowest visit category, we move to the patient category with the next lowest visit rate and transfer a patient to the physician with the least overflow. This process of transferring patients is continued until the difference between maximum overflow of the practice and the reference overflow is small enough. We call this Heuristic 1, or H1.

Notice that in H1, we may have to shift a very large number of patients from low visit rate categories to achieve identical overflows in the practice. This may not be a bad strategy since relatively healthy patients have a lower chance of having formed a strong bond with the PCPs and are therefore more likely to change their PCPs.

In Heuristic 2, or H2, a different approach which involves all patient categories in the patient transfers is analyzed. As before we start with the current panel design and identify the physicians with the highest and lowest overflow values. We then transfer one patient from the patient category with the lowest visit rate to begin with, update the overflow values of the two physicians and again identify the physicians with the highest and lowest overflow values. If the current value of maximum overflow and the reference overflow is still large, we switch – in contrast to H1 – a patient from the category with the next lowest visit rate. Thus we move from one category to the next, whereas in Heuristic 1, we tried to exhaust all possibilities in the lowest visit category. In Heuristic 2, patients are more evenly moved across the different categories, but more importantly *fewer* patients are moved in relation to Heuristic

1. The downside is that patients with chronic conditions who are more likely to have a strong relationship with their PCP will also be transferred in Heuristic 2.

While H1 and H2 lie at two ends of the spectrum, a practice manager can be more creative in his transfer choices. Patient and physician surveys as well as past visit patterns can be used to make more intelligent transfer choices that minimize disruption. In practice, patient reassignment is a dynamic process, which will be carried out over a period of time, as new patients are empanelled in the practice, when physicians leave or retire (thus leaving their panel to be reassigned among still working physicians). In addition, practices can use surveys to determine the willingness of patients to change their PCPs, thus creating a pool of patients who are amenable to changing their PCPs.

2.8 Case Study

2.8.1 Data description

We use data from the Primary Care Internal Medicine (PCIM) practice at the Mayo Clinic in Rochester, MN. This practice empanels around 27,000 patients and employs 39 physicians. Many of these physicians worked part time. Panel data enabled us to identify which patients belonged to which physician. Patient level data included the number and type of chronic conditions afflicting each patient as well as the number of visits for each patient for 3 years (2004, 2005 and 2006). The list of chronic conditions included commonly occurring diseases such as hypertension, depression, diabetes, osteoporosis, urinary tract infections, hyperlipidemia, coronary artery disease and otitis. As discussed before, the number of comorbidities are used to come up with

patient categories and this gives us 8 patient categories in all. To determine the p_i values for each comorbidity count, we first determine A_i, which is the total number of appointment visits for all patients with i comorbidities in the population for a long period of time, say a year. If N_i denotes all patients with i comorbidities, and if there are T workdays in a year, then:

$$p_i = \frac{A_i}{N_i * T}. \tag{2.15}$$

Assuming there are 250 workdays in a typical year, we are now able to calculate the per day request probability p_i for each patient category. The method is similar to the one proposed in Green et al. [2007a]. The values are listed in the Table 2.3 below. It is also possible to calculate the p value for the entire population. If A is the total

Table 2.3: Binomial p_i values for each patient category

p_0	p_1	p_2	p_3	p_4	p_5	p_6	p_7
0.0062	0.0106	0.0149	0.0199	0.0260	0.0298	0.0380	0.0412

visits generated by the total population of N patients, then:

$$p = \frac{A}{N * T} = 0.0143 \tag{2.16}$$

This value will be used to set the capacity of physicians in the test practices created based on our data. The idea is to replicate the default process by which practices typically assign capacity – they recognize that capacity should increase with panel

size, but generally do not consider case-mix in how they determine capacity. Thus, if a physician's panel size is L_j, then the physician's capacity, C_j, is assigned as follows:

$$C_j = \lceil (L_j * p + 0.1 * L_j * p) \rceil \tag{2.17}$$

The physician is given 10% more slots than the mean demand $L_j * p$. Setting it equal to the mean – as many practices might, since they remain unaware of the impact of variance – would mean that each physician's utilization would be 100%, leading to an unsustainable system. The 10% additional slots ensure that there are a few extra slots to buffer variability in demand. Yet the utilization of the physician will still be sufficiently close to 100%, as it is for most PCPs practicing in the U.S. today. The above expression rounds up to the closest values, since the number of appointment slots per physician per day is typically an integer. We note that our approach can work with any other capacity inputs as well.

Our goal is not to obtain results specific to Mayo Clinic data. Rather it is to use the data to generate a series of "test" practices with 2 and 4 physicians, with different case-mixes to illustrate the impact of case-mix and our heuristics. The majority of practices in the U.S. have 5 physicians or less, so our practice sizes are appropriate. Furthermore, larger practices tend to be divided into smaller self-contained subgroups to ensure continuity. We note, however, that our method is not computationally constrained in any way and can address larger practices as well.

2.8.2 Panel redesign for test practices

Tables 2.4, 2.5, 2.6 and 2.7 provide detailed results for our 4 test practices. The table format allows a reader to see the panels, case mixes and corresponding measures clearly. We consider the equal and unequal capacity case and under each we test a 2 physician case and a 4 physician case. In the first two test practices, the physicians have approximately the same panel sizes and hence the same capacity. In the next two, physicians have different panel sizes and hence have different capacities. The capacities are calculated as described above, based on panel size only. The physicians are numbered based on the original Mayo Clinic data (which had 39 physicians) to distinguish them from each other. We note that any combination of the 39 physicians from the data set can be considered in a similar way.

In the tables, we present panel case mixes before and after redesign, the corresponding means and variances for each panel, the overflow and the utilization for each physician. We also present panels designed based on the 1) Capacity Ratio 2) Heuristic 1 and 3) Heuristic 2. Note that the capacity ratio rule allocates patients from each category i to each physician j as follows: $x_{ij} = (C_j/C) * (N_i)$, where $C = \sum_{j=1}^{J} C_j$ is total capacity of the clinic. In the equal capacity case, when $C_1 = C_2 = ... = C_J$, the allocation reduces to $x_{ij} = (N_i/J)$, which gives the optimal solution (see Section 2.6). In the unequal capacity cases, $x_{ij} = (C_j/C) * N_i$ is a heuristic that is expected to perform well, but will not necessarily be optimal. For these cases, reference overflow values are used as the benchmark for comparisons.

In both Heuristic 1 and Heuristic 2, we start with the current panels or current case-mix and switch patients (as described in the previous section) until the required

maximum overflow value is reached. In each heuristic (including the Capacity Ratio), we list the number of patients switched from each comorbidity group as well as the total number of patients switched.

Table 2.4: Results for Test Practice 1: 2 physicians with equal capacity. O_{ref} for this practice is 0.24. The number of patients switched is provided as a separate row under each heuristic. The total patients switched by a particular heuristic appears under the Panel Size column.

		Comorbidity Count												
		0	1	2	3	4	5	6	7	Panel Size	μ_j	σ_j^2	C_j	O_j	Utilization
Current	Phy 4	33	23	60	32	42	70	14	40	205	1495	21.99	21.58	24	0.33	0.92
	Phy 28	41	83	85	29	92	11	11	32	102	1469	19.64	19.31	24	0.16	0.82
	# Switched	0	0	0	0	0	0	0	0		0					
Capacity Ratio	Phy 4	37	53	72	31	22	40	12	8	36	153	1481	20.80	20.43	24	0.24	0.87
	Phy 28	37	53	73	31	12	41	12	7	36	154	1482	20.83	20.46	24	0.24	0.87
	# Switched	43	12	12	30	16	4	5	2		124					
Heuristic 1	Phy 4	144	36	60	32	42	70	14	40	205	1307	20.81	20.42	24	0.24	0.87
	Phy 28	606	83	85	29	92	11	11	32	102	1656	20.81	20.47	24	0.24	0.87
	# Switched	188	0	0	0	0	0	0	0		188					
Heuristic 2	Phy 4	32	53	53	31	72	63	13	73	3	140	1442	20.80	20.43	24	0.24	0.87
	Phy 28	42	53	92	30	62	18	11	83	9	167	1521	20.83	20.46	24	0.24	0.87
	# Switched	7	7	7	7	7	7	6	5		53					

In **Test Practice 1** shown in Table 2.4, while the two physicians have almost the same panel size and therefore the same capacity (24), differences in their case-mix result in significantly different overflow values. Physician 4 would therefore be unable to provide timely access and continuity to her patients. It is quite likely that the patients of Physician 4 that are unable to secure an appointment would end up seeing Physician 28. When the panels are redesigned, their overflow values can be

made even. Physician 28's overflow and utilization increase as she receives some of Physician 4's patients.

The Capacity Ratio heuristic which is optimal for this practice evens the case-mix differences between the physicians and in the end results in similar panel sizes as before. However, in order for the two physicians to achieve the allocation suggested by Capacity Ratio, 124 patients need to be switched – this includes a number of high comorbidity patients. Heuristic 1 achieves identical overflows by starting with the original case-mix and then transferring 0 comorbidity (healthy) patients from Physician 4 to Physician 28. As mentioned before, these patients are more likely to accept a PCP change. Notice that Heuristic 1 results in very different panel sizes as a result. Heuristic 2, on the other hand, switches patients evenly across categories but this does mean that higher comorbidity patients will be switched. The total patients switched however is only 53, about half of what Heuristic 1 requires. The panel sizes are different after Heuristic 2, but the difference is not as drastic as that produced by Heuristic 1.

For **Test Practice 2** (Table 2.5), all four physicians have a capacity of 17 and approximately the same panel size. These are the same four physicians whom we used to motivate the paper in Section 2.4. We see here too Physicians 34 and 8 have significantly higher overflow. The Capacity Ratio heuristic evens out the differences but this comes at a cost of shifting 193 patients. Heuristic 1 switches 229 patients, which constitutes 5% of the total patients, but all of them are 0 comorbidity patients. Heuristic 2 switches only 62 patients (only 1.5% of the total patients) but this does include a few high comorbidity patients. The difference in the number of patients

switched (from each patient category and in total) can clearly be observed from Table 2.5. Notice that both Heuristic 1 and Heuristic 2 are able to reach the overflow values that the Capacity Ratio allocation produces, which is optimal in this equal capacity case. Both the capacity ratio algorithm and the heuristics are able to balance the utilization and overflow frequency.

In **Test Practice 3** (Table 2.6), Physician 20 has more patients in her panel and also has more capacity (21) compared to Physician 24 (15). However, the former's overflow is more than double the latter's. There is a clear case for panel redesign here, since Physician 20's current capacity of 21 slots per day is already quite high and mostly likely cannot be increased anymore. This is especially true since PCPs are responsible for numerous other non-visit tasks during the day, such as attending phone calls, coordinating with specialists her patient might have recently visited and so on. The Capacity Ratio reduces the imbalance in panel workloads somewhat but clearly does not provide the optimal solution. Notice that the utilizations (which are calculated using the mean demands and the capacity of the physician) are perfectly balanced under Capacity Ratio, but the overflows are not. This is because the utilization (μ_j/C_j) does not consider variance but the overflow frequency does. Moreover Capacity Ratio switches 142 patients. Heuristic 1 and 2, on the other hand, produce overflows that are almost identical to the reference overflow (0.264). Heuristic 1 switches 172 healthy patients, while Heuristic 2 switches 52 patients in total from all the categories. Thus with regard to both overflow and patients switched, the H1 and H2 are better than Capacity Ratio.

Table 2.6: Results for Test Practice 3: 2 physicians with unequal capacities. The reference overflow value, O_{ref} for this practice is 0.264. The number of patients switched is provided as a separate row under each heuristic. The total patients switched by a particular heuristic appears under the Panel Size column.

		Comorbidity Count													
		0	1	2	3	4	5	6	7	Panel Size	μ_j	σ_j^2	C_j	O_j	Utilization
Current	Phy 20	255	314	289	223	124	54	21	1	1281	19.33	18.97	21	0.35	0.92
	Phy 24	255	262	189	107	52	25	5	1	896	11.64	11.45	15	0.16	0.78
	# Switched	0	0	0	0	0	0	0	0	0					
Capacity Ratio	Phy 20	297	336	278	192	102	46	15	1	1267	18.01	17.69	21	0.24	0.86
	Phy 24	213	240	200	138	74	33	11	1	910	12.96	12.73	15	0.28	0.86
	# Switched	42	22	11	31	22	8	6	0	142					
Heuristic 1	Phy 20	83	314	289	223	124	54	21	1	1109	18.26	17.90	21	0.26	0.87
	Phy 24	427	262	189	107	52	25	5	1	1068	12.71	12.51	15	0.26	0.85
	# Switched	172	0	0	0	0	0	0	0	172					
Heuristic 2	Phy 20	247	306	282	216	117	47	14	0	1229	18.26	17.92	21	0.26	0.87
	Phy 24	263	270	196	114	59	32	12	2	948	12.71	12.50	15	0.26	0.85
	# Switched	8	8	7	7	7	7	7	1	52					

Finally as can be seen from Table 2.7, in **Test Practice 4**, there are four physicians with different panel sizes and capacity values (24, 17, 15 and 14 respectively). Notice, however, that the overflow and utilization values are not dramatically different to begin with (at least in relation to Test Practice 3). In this case, the practice may decide that no redesign is required. We note here that our approach and presentation of performance measures will help practices come to such a conclusion.

Table 2.7: Results for Test Practice 4: 4 Physicians with unequal capacities. The reference overflow value, O_{ref} for this practice is 0.177. The number of patients switched is provided as a separate row under each heuristic. The total patients switched by a particular heuristic appears under the Panel Size column.

		Comorbidity Count												
		0	1	2	3	4	5	6	7	Panel Size	μ_j	σ_j^2	C_j	O_j	Utilization
Current	Phy 28	418	385	299	211	111	32	10	2	1469	19.64	19.31	24	0.16	0.82
	Phy 19	299	293	212	147	77	26	6	1	1062	14.10	13.86	17	0.22	0.83
	Phy 17	274	245	189	98	52	23	11	1	894	11.57	11.37	15	0.15	0.77
	Phy 12	244	233	162	107	46	27	9	2	830	10.96	10.77	14	0.18	0.78
	# Switched	0	0	0	0	0	0	0	0	0					
Capacity Ratio	Phy 28	426	399	297	194	98	37	12	2	1465	19.36	19.03	24	0.14	0.81
	Phy 19	310	290	216	142	73	28	10	2	1071	14.24	14.00	17	0.23	0.83
	Phy 17	259	242	181	118	60	22	7	1	890	11.75	11.55	15	0.17	0.78
	Phy 12	240	225	168	109	55	21	7	1	826	10.91	10.73	14	0.17	0.78
	# Switched	28	33	11	27	15	8	6	1	129					
Heuristic 1	Phy 28	458	385	299	211	111	32	10	2	1508	19.89	19.56	24	0.18	0.83
	Phy 19	219	293	212	147	77	26	6	1	981	13.60	13.37	17	0.18	0.80
	Phy 17	315	245	189	98	52	23	11	1	934	11.82	11.63	15	0.18	0.79
	Phy 12	243	233	162	107	46	27	9	2	829	10.95	10.77	14	0.18	0.78
	# Switched	81	0	0	0	0	0	0	0	81					
Heuristic 2	Phy 28	418	387	301	213	112	34	11	3	1479	19.89	19.56	24	0.18	0.83
	Phy 19	295	290	209	144	74	23	3	0	1038	13.61	13.39	17	0.18	0.80
	Phy 17	278	246	190	99	54	24	13	1	905	11.79	11.60	15	0.17	0.79
	Phy 12	244	233	162	107	46	27	9	2	830	10.96	10.77	14	0.18	0.78
	# Switched	4	3	3	3	3	3	3	1	23					

As in Test Practice 3, we note that Capacity Ratio is a good heuristic and reduces the imbalance between physicians but does not give the optimal overflow. It also requires that 129 patients be moved, despite the fact that overflow differences between the physicians are not significant. Heuristic 1 and 2 are more effective in reducing the overflow, but also move fewer patients compared to Capacity Ratio. Heuristic

1 switches 2%, whereas Heuristic 2 only changes 0.5% of the total patients. As before this is because, Heuristic 1 affects only the healthy patients, while Heuristic 2 involves patients from all categories.

2.8.3 Quantifying the price of continuity

We can also measure the price of continuity by quantifying the gap between O_{ref} and O_{CP} in terms of the number of new patients who can can be empanelled. Recall that O_{ref} is a surrogate for the best possible access that the physicians in the practice can provide with the available capacity once panels are redesigned and assuming that physicians do not see each others patients. In contrast, O_{CP} is the overflow of a practice in which all panel demand is aggregated and all physician capacity is pooled. The latter provides improved access to care (lower wait times) but at the expense of continuity. (Since $O_{CP} \leq O_{ref}$ from Sec. 2.6)

Table 2.8: Price of continuity in terms of number of patients, where TP: Test Practice

Capacity	**TP**	**# physicians**	**Capacity**	**Ocp %**	**Oref %**	**Patients added**
Equal	1	2	48	16	24	127
	2	4	68	17	31	255
Unequal	3	2	36	18	26	100
	4	4	70	3	18	500

If a practice cares more about access to care than continuity, then how many patients could it have added if O_{CP} is allowed to increase and until it equals O_{ref}? In other words, if the access performance as measured by overflow frequency is held constant, how many more patients can a pooled practice with no concept of continuity empanel compared to a dedicated practice where patients only see their own PCP?

We quantify the number of patients that could be empaneled for each of the test practices in Sec. 8.2.

In Test Practice 1, which consists of 2 equal capacity physicians, 127 new patients could have been empanelled if the current O_{CP} value of 0.16 is allowed to increase to the O_{ref} value of 0.24. For this calculation, we assume that the new patients added have the same comorbidity mix that the practice currently has. For example, 750 of the 2963 total patients (around 25%) in Test Practice 1 were 0 comorbidity patients. Since this may be a fair reflection of the demographics of the neighborhood in which the practice is located, we assume that 25% of the 127 new patients that the practice can empanel will also be 0-comorbidity patients. Similar calculations apply for other comorbidity counts.

The addition of new patients implies a loss of continuity since any physician in the practice can see any patient. There is no single PCP who coordinates the patients' care. In a fee-for-service system, where physicians are reimbursed based on the number of visits, the revenues for the practice will increase as will the overall ability to access physicians, but patient centeredness and possibly physician satisfaction will likely to decrease.

Among the two 4-physician practices, Test Practice 2 will be able to add 255 patients at the expense of continuity while Test Practice 4 will be able to add 500 patients at the expense of continuity. This difference is because of two reasons. First, comorbidity counts are higher in Test Practice 2 compared to Test Practice 4. Second, utilization and overflow values are lower in Test Practice 4 to begin with, allowing for a greater number of patients to be added.

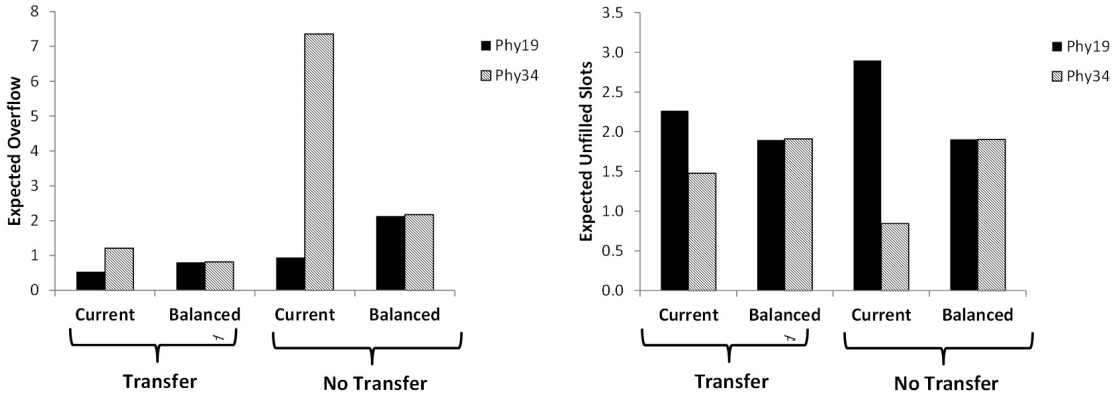
Thus our framework allows a practice to look at extremes of best possible continuity and best possible access and make their empanelment decisions accordingly.

2.8.4 Impact on other measures

We have so far investigated overflow frequency and utilization. We now look at Expected overflow (EO) and Expected unfilled slots (EU). Expected overflow, which was explained in Section 2.4, represents the average number of patients who were not able to get appointments. Expected unfilled slots tells us how under-utilized each physician is. To test the impact on these two measures, we choose Physicians 19 and 34, from Test Practice 2. Both these physicians have equal capacity (17) and before their panels are redesigned, their overflow frequencies were 0.22 and 0.42 respectively. We calculate EO and EU for both physicians before redesign (Current) and after redesign (Balanced). The heuristic used for redesign is Capacity-ratio, which gives an optimal allocation since the two physicians have the same capacity.

Since there is no closed form expressions for EO and EU, we simulate 10,000 realizations of demand, sampled from the binomial distributions appropriate for each patient category. Each realization represents a day in the model. If the physicians have any backlog it is transferred to the next day. We also investigate the impact of *sharing or transferring* patients. That is if a physician has capacity available after seeing her own patients, then she is allowed to see the other physician's patients (if the other physician has a backlog), at the expense of continuity. We compare this case against the dedicated case, where the physicians do not share or transfer their patients; that is, they maintain continuity at the expense of timely access.

Figure 2.3 clearly shows the benefits of redesign (Balanced versus Current). The benefits are especially significant when the two physicians do not share their patients (the No Transfer case). If the physicians are not allowed to transfer patients and case mixes remain the same then the resulting expected overflow is almost unsustainable (for Physician 34 especially), resulting in poor access. Panel redesign produces more even EO profiles when sharing is allowed (Transfer case), but the difference is not as significant as in the no-transfer case. We notice here that sharing of patients mitigates the poor timely access problem. The unevenness in expected unfilled slots between physicians is leveled with the balanced case mixes.



(a) Expected Overflow (b) Expected Unfilled Slots

Figure 2.3: Results for 2 physicians with equal capacity

These results suggest that even if the practices are unwilling to redesign panels, sharing of patients between physicians is a viable alternative, especially if a practice consists of 2-3 physicians. Moreover the sharing can be restricted to same-day requests for which continuity is not always necessary or desired by the patients. While this is not the ideal scenario, access is improved at the cost of continuity of care. If

the physicians are keen on providing continuity then it is clear that the panels have to be redesigned. We find similar results while testing other pairs of physicians, but in the interest of keeping the paper concise these results are not presented.

Table 2.5: Results for Test Practice 2: 4 physicians with equal capacity. O_{ref} for this practice is 0.31. The number of patients switched is provided as a separate row under each heuristic. The total patients switched by a particular heuristic appears under the Panel Size column.

		Comorbidity count													
		0	1	2	3	4	5	6	7	Panel Size	μ_j	σ_j^2	C_j	O_j	Utilization
Current	Phy 39	290	296	218	145	84	27	12	5	1077	14.73	14.47	17	0.28	0.87
	Phy 8	260	249	226	161	108	42	14	3	1063	15.54	15.26	17	0.35	0.91
	Phy 19	299	293	212	147	77	26	6	1	1062	14.10	13.86	17	0.22	0.83
	Phy 34	214	253	223	177	115	44	21	5	1053	16.16	15.85	17	0.42	0.95
	# Switched	0	0	0	0	0	0	0	0	0					
Capacity based	Phy 39	266	272	220	157	96	34	14	3	1062	15.11	14.83	17	0.31	0.89
	Phy 8	266	272	220	157	96	34	14	3	1062	15.11	14.83	17	0.31	0.89
	Phy 19	265	273	219	158	96	35	13	4	1063	15.15	14.87	17	0.32	0.89
	Phy 34	266	274	220	158	96	36	12	4	1066	15.17	14.90	17	0.32	0.89
	# Switched	58	44	9	23	31	16	9	3	193					
Heuristic 1	Phy 39	357	296	218	145	84	27	12	5	1144	15.14	14.89	17	0.32	0.89
	Phy 8	194	249	226	161	108	42	14	3	997	15.13	14.85	17	0.31	0.89
	Phy 19	461	293	212	147	77	26	6	1	1223	15.11	14.87	17	0.31	0.89
	Phy 34	51	253	223	177	115	44	21	5	889	15.15	14.84	17	0.32	0.89
	# Switched	229	0	0	0	0	0	0	0	229					
Heuristic 2	Phy 39	292	298	220	147	86	30	14	7	1094	15.13	14.86	17	0.31	0.89
	Phy 8	258	247	224	159	106	39	12	1	1046	15.14	14.87	17	0.31	0.89
	Phy 19	305	299	218	153	83	31	11	6	1106	15.11	14.84	17	0.31	0.89
	Phy 34	208	247	217	171	109	39	16	0	1007	15.15	14.87	17	0.32	0.89
	# Switched	8	8	8	8	8	8	7	7	62					

2.9 Conclusions and Implications for Practice

In summary, we have shown that case-mix is an important consideration in primary care. Physicians with the same panel size but different case-mixes can have very different overflow frequencies. We have characterized how overflow frequencies can vary from physician to physician and demonstrated, using actual data from a primary care practice, how these imbalances in supply and demand can be minimized in the long term.

To implement our results, a practice will have to collect appointment request rates of its patient population from historical data. Two to three years worth of visit data should be sufficient to classify patients according to their visit patterns. With the increasing use of electronic records, such data should be easily available. Practices can use the opportunity to update information about currently active patients and obtain more precise information about panel sizes.

Once this assessment is complete, practices can then begin to benchmark their current performance by comparing the overflow frequencies of the physicians in relation to one another and in relation to the reference overflow derived in this paper. Panel redesign options can be easily tested, in a manner similar to Tables 2.4, 2.5, 2.6, 2.7 and the least disruptive options of redesigning panels can be identified. In general clinics should be aware that O_{ref} values of 0.3 or above, which result in high utilization, should be avoided.

All overflow frequency calculations derived in this paper can be easily carried out in an Excel spreadsheet. The American Academy of Family Physicians (AAFP) has an Excel spreadsheet tool for panel size calculations (Murray et al. [2007]). However,

it uses only the mean, does not consider case-mix and does not consider the impact of variance. The Excel tool provided by Green et al. [2007a] allows practices to decide on panel size for a single physician based on overflow frequency. The impact of variance is also considered in their calculations. The results in this paper extends the Green et al. [2007a] framework, to allow for an Excel tool that 1) quantifies the impact of case-mix; 2) calculates a benchmark overflow value for a group practice; and 3) allows for testing of various panel redesign options in the long term. A preliminary version of our Excel spreadsheet is available for free at people.umass.edu/hbalasub/PanelDesignSpreadsheet.xlsx.

Our model does have limitations, which provide opportunities for future investigation and model refinement. We do not consider seasonality and day of week effects on overflow frequencies. In Savin [2006] (Section 3.2.7), he analyzes the effect of seasonality and day-to-day variability in a primary care practice and observes that variations can be quite high. To model this effect, he adjusts the probability that a patient requests an appointment for a specific day or month. Our category specific p_i values can also be adjusted depending on the time of the year or day of the week. Savin [2006] suggests that to cope with such variations practices will either have to adjust panel sizes, or flexibly adjust the capacities of the physicians. In addition, practices can leverage the benefits of working in groups – an aspect we consider in this paper. In peak seasons or busy days, urgent same-day requests could be flexibly shared by a small group of two-three physicians. As Section 2.8.4 shows, such flexibility can improve access; the compromises in continuity will be small so long as provider team is small. For more details on how same-day flexibility can be de-

signed to balance access and continuity under different utilization levels, we point to Balasubramanian et al. [2013] and Balasubramanian et al. [2011].

Another extension worth considering is whether physician practice style has an impact on visit rates and consequently overflow frequencies. This can happen if some physicians schedule more follow-up visits than others on average. Recall that in the current model, demand is controlled by the p_i values for the comorbidity categories, which in turn is decided by the total number of visits from each category over a long period (2-3 years). Now, as an example, if we were able to determine – through new empirical data and appropriate statistical tests – that physician j scheduled twice as many visits for high comorbidity count patients compared to physician k, then the p_i values for that category would accordingly have to be physician specific. So not only do higher comorbidity patients have higher visit rates (which is indeed the case and is the premise of our paper), but some physicians schedule more visits for these patients than others, with implications for the overflow frequency. This is an interesting direction for future work, and would require careful collection of new physician-specific appointment data.

As mentioned earlier, our modeling approach is designed for aggregate level panel management decisions. While we do not explicitly consider different appointment types, such as prescheduled and same-day, a high overflow frequency will be correlated with the inability to provide access for both types of appointments. In the same way, although no-shows are not a part of our model, well designed panels can only reduce the impact of no-shows, by improving time to earliest available appointments. See Green and Savin [2008] for a discussion. Finally, patients with more comorbidi-

ties are more likely to have longer appointments than healthy patients. Our values for overflow frequency are therefore likely to be slightly smaller than those found in practice. However, in a relative sense, our approach will still correctly identify the imbalances in supply and demand across physicians. If anything, redesign will have an even greater effect.

CHAPTER 3

PRIMARY CARE PRACTICE DESIGN UNDER CASE-MIX: JOINT CONSIDERATION OF ACCESS TO CARE AND CONTINUITY OF CARE

3.1 Introduction

Primary care can prevent illness, improve health outcomes and reduce mortality (Starfield et al. [2005]). Providing communities with high-quality primary care is set as priority in many countries healthcare agenda. To build a successful primary care delivery system, access to care and continuity of care are two crucial cornerstones.

The concept access to care has a broad meaning (Aday and Andersen [1974]). Some researchers equate it to the availability of health system resources in an area, while others relate it to characteristics of the population, e.g., incomes, insurance coverage and attitudes toward medical care. Simply put, accessibility to primary care can be thought of as how easy it is for a patient to receive primary care when

he/she needs it. Previous research has developed quantitative measures of accessibility, among which one of the most important measures is the appointment delay (Balasubramanian et al. [2010]). The appointment delay refers to the time between a patient's call for an appointment and her actual appointment date. The shorter the appointment delay, the earlier a patient can receive the medical service, and hence the more accessible the primary care service is.

The other crucial cornerstone for a successful primary care system is continuity of care. Saultz [2003] summarizes this concept in a hierarchical way: 1) informational continuity means patient information is transferred when she sees another provider; 2) longitudinal continuity of care refers to patients receiving most of their care from the same provider; 3) interpersonal continuity implies an ongoing relationship and trust existing between each patient and a personal physician. The most commonly-used concept for continuity of care is the longitudinal continuity of care, which is usually defined as the percentage of time that the patient is seen by her own primary care provider (Bice and Boxerman [1977]).

Ideally, a primary care practice would like to improve both access to care and continuity of care offered to its patients, but these two goals are often conflicting (2). For example, many primary clinics aim to improve access to care and reduce appointment delays by implementing open access (Murray and Tantau [2000]). In doing so, they try to provide a majority, if not all of, the patients with same-day appointments. To build up enough service capacity, they may choose to form practice teams with multiple providers, say two to three, sharing their patients. Though this pooling strategy does improve service capacity, it may lead to loss of continuity of

care, because there is no guarantee that patients will always be seen by their own providers unless the patients' situations dictate that or their own providers happen to have open slots during their visit. Indeed, among those Open Access trials that failed, many are because the loss of continuity as a price to pay for speedy access is just too high (Phan and Brown [2009]).

In the design of a primary care practice, case-mix is another crucial factor that needs to be accounted for. Case-mix refers to the type of patients served by a practice. Because different types of patients may have different visit frequencies as well as various demand for providers' consultation time, case-mix directly influences the "demand" side of a primary care practice. For example, Potts et al. [2011] have calculated the disease burden of a physician's panel by using the risk categories set for chronic diagnoses in order to decide on the support the physician needs from nurse practitioners (NPs). The goal of this paper is to develop methodologies to quantify and evaluate access to care and continuity of care in primary care practices, taking into account the impact of case-mixes.

Adding more patients in a physician's panel increases the physician's workload, and thus leads to longer appointment delays. The panel size, as explained in Chapter 2, is the number of patients that a physician (group) is held accountable for (Murray et al. [2007]). Given the same panel size, a physician's workload is larger if patient acuity level is higher because patients visit the clinic more often and each visit might also take longer time (Knox and Britt [2004]; Roos et al. [1998]). To reduce appointment delays and improve practice, there are two major operational strategies. One is to take the advantage of economies of scale by forming a practice team and

pooling service capacity together, but there would be a loss of continuity of care. Another strategy is via panel redesign, i.e., to reallocate patients to providers' panels according to patient needs and individual provider service capacity so that the whole care team is used more effectively (Balasubramanian et al. [2010]); but reallocating patients may not be an easy task as it takes time and effort, and involves changing existing patient-PCP relationships. The qualitative effect of these two strategies is clear. The question, however, is how to quantify these effects ex-ante and also adjusted for case-mixes.

In this chapter, we will use queueing theory to develop methods that enable us to conduct such quantitative analysis, which should provide useful information for practice change. Queueing theory concerns the study of wait lines (Gross and Harris [1985]). It can translate customer arrival characteristics and service patterns into measures of waiting experienced by the customers, e.g., average waiting time and the chance that customers will be delayed in the service process. In this paper, we measure access to care by appointment delays (i.e., wait time) and operationalize continuity of care by the percentage of patients who see their own primary care providers. Since we are interested in studying the relationship among panel size (which, to be discussed shortly, is directly related to patient appointment demand), provider service capacity and patient appointment delays, queueing theory is an ideal tool.

We consider three typical practice designs used in primary care. The first design is a dedicated service model where patients only see their own providers. This design can also be viewed as a solo-practitioner service where the provider serves her

own patients only. The second and third designs are both group practice models involving multiple providers working in the same team. The difference is that, in the second design, some patients have their dedicated providers while others see anyone available. In the third design, patients see any available provider. For each of these designs, we develop corresponding queuing models and derive performance measures for both access to care and continuity of care. We use data collected from the Mayo Clinic to populate our models and discuss how these measures change among designs. All these measures can be computed via closed-form formulas, and they can be easily evaluated using spreadsheet tools like Excel or even just calculators.

With the recent passage of the Patient Protection and Affordable Care Act, more than 30 million Americans are expected to gain healthcare coverage in the U.S. However, many areas in the nation are facing severe shortage in primary care workforce (HHS [2009]). Compounding the increase in patient volumes and the shortage of primary care workforce is the aging population and the epidemic of chronic diseases, which will likely give rise to more patients with multiple comorbidities requiring more physician time and resources. To reform primary care delivery in the U.S., many practices are engaged in transforming into Patient Centered Medical Homes (Nielsen et al. [2012]), one of the most important objectives being to form a coordinated and integrated care team that provides patient centered care. Yet, there is a lack of scientific and systematic methods that can inform the formation of such teams and the allocation of workload among different team members to achieve the best outcome. Our study provides a tool to assess the supply demand dynamics, conduct capacity planning and inform practice design for primary care teams.

3.2 Methods

3.2.1 The models

We use queueing models to describe the operations in different primary care practice designs, using the formulas in the Appendix A. As an example, consider a single physician's practice. Patients in the physician's panel call and request an appointment. To better understand our model, suppose for now that patients will take the earliest appointment slot available and the service time for each patient is deterministic with a common length (we will relax this assumption later). Thus the provider knows exactly when to schedule this patient upon her request. In particular, incoming appointment requests are registered on the provider's work schedule in the order they arrive. The provider's schedule is the queue in our models. The queue here is not the physical waiting line of patients in the clinic, but rather a virtual list for those who have not yet been seen by the provider.

During office hours, the provider sees patients and shortens the queue. When the practice is closed, no one joins or leaves the schedule, i.e., the queue remains intact. If we remove the non-office hours from the time horizon, we can view the provider's work schedule as a continuous queueing process, where jumps and drops in this queue correspond to the arrival of an appointment request and the service completion of a patient, respectively.

In reality, patient preferences, punctuality and type of appointments (prescheduled versus same-day) may play an important role in practice operations. These factors can be considered by more sophisticated frameworks, e.g., Wang and Gupta [2011], which usually focus on intra-day operations; while our goal is to evaluate the

access to care and continuity of care in primary care practices across days. To that extent, our analysis is on a more strategic level, and thus we choose not to incorporate too many intra-day scheduling details in our models. Omitting these details leads to much more accessible formulation that provides quantifiable outcome measures for practical use. More importantly, several recent studies support the use of such models in setups like ours (Green and Savin [2008]; Liu and D'Aunno [2012]). In particular, by comparing with more realistic simulation models that consider patient preference and other scheduling details, Green and Savin [2008] show that queueing models can yield relatively accurate estimates for panel sizes.

One interesting and innovative feature of our models is that they can account for case-mixes. Case-mix refers to the type of patients in a panel, and it can be characterized by various attributes, such as age, gender and the chronic conditions afflicting the patient (Balasubramanian et al. [2010]). The idea is to group patients into "categories," and within each category patients have similar demand pattern and needs for providers' time and resources. Using data from the Mayo Clinic, we will discuss how to categorize patients shortly.

In Figure 3.1, the Greek letters λ and μ represent the patient arrival rate and provider service rate, respectively. We now describe the specifics of our models. The appointment rate of a patient is assumed to follow a Poisson Process with a rate λ_i^0 per day, for patient category i. If there are N_i patients in category i, then the appointment rate from this category is $\lambda_i = N_i * \lambda_i^0$. If there are M patient categories, then the panel size is the number of patients in all categories, i.e., $N_1 + N_2 + ... + N_M$; and the joint arrival process is also a Poisson process whose arrival rate is the sum of

those of its constituting arrival streams, i.e., $\lambda_1 + \lambda_2 + ... + \lambda_M$. The Poisson process is a widely used customer arrival model (Gross and Harris [1985]). It is especially reasonable in our outpatient primary care setting as patients requests usually arrive one at a time, and they can be treated independent of one another.

Since patients in different categories may require different amounts of service time, the service time of a physician also needs to be adjusted for the case-mix. For instance, a physician with more of higher acuity level patients in her panel should have a lower number of appointments per day to accommodate for longer service times. To adjust for case-mix, we calculate the average appointment duration for a physician by taking the weighted average of the service times from different categories of patients, where the weights correspond to proportions of the arrival rate from each patient category. Thus the weighted average service time μ is calculated as $\sum_i^N \frac{\lambda_i}{\sum_i^M \lambda_i} \mu_i$, where μ_i is the service time for category i.

With the above model description in mind, we proceed to discussing the three practice designs (Figure 3.1) we will investigate in this article.

The first design is a dedicated service model where patients always see their own provider. This design can also be used in a multi-provider practice, where each provider practices as an independent single physician. The second design is a group service model with partial pooling of provider service capacity, where some patients have dedicated providers while others are flexible. In particular, dedicated patients to provider 1 have an arrival rate of λ_1 and they will wait as long as provider 1 is busy. Similarly, dedicated patients to provider 2 arrive at rate λ_3 and they will wait as long as provider 2 is busy. Another stream of patients arriving at rate λ_2

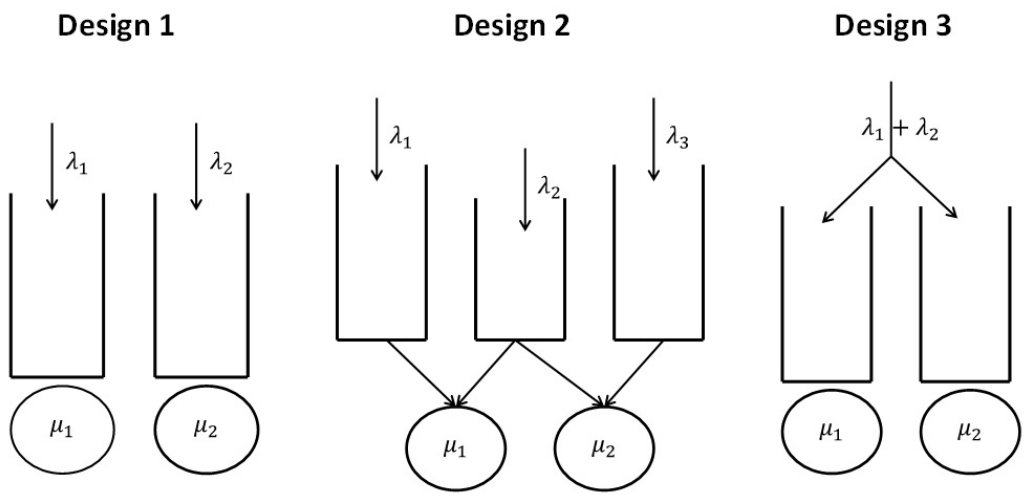


Figure 3.1: Practice designs

are flexible patients; they will see any available provider and they wait only if both providers are busy. The third design is a group service model with complete provider capacity pooling, where patients will see any provider who is available.

Finally, we relax our service time assumption in our analysis. Recall that when describing our models, we suppose that the provider service time is deterministic. Under this service time assumption, we typically do not have closed-form expressions for the performance measures that we are interested in. For better tractability, we relax this assumption and assume that service times are random, which, in particular, follow exponential distributions. On the one hand, random service times bring some variability into the service process and seem to resemble practical settings better. On the other hand, queues with exponential service times are usually easier to analyze and often have closed-form expressions for their performance measures. Furthermore,

previous studies show that variability in provider service time typically does not have a significant impact on the productivity of a practice, while the mean service time is a more important determinant (Liu and D'Aunno [2012]; Liu et al. [2012]). For all reasons above, we will focus our analysis on queues with exponentially distributed service times.

3.2.2 Data and model parameters

We analyze the patient population (around 20,000 patients) empanelled at the Primary Care Internal Medicine Practice (PCIM) of the Mayo Clinic in Rochester, Minnesota. Our data constitute patient visits over three years 2003-06 to 39 physicians at PCIM. Detailed analysis on patient demand rates has been reported in an earlier study by 2. We recapitulate the key results here for convenience. Their analysis reveals that comorbidity count (CC) is the strongest predictor for patient demand rate. Thus we divide patients based on the number of comorbidities they had. In all, there are 8 patient categories as patients with more than 7 comorbidities were extremely rare. Our categorization is consistent with earlier literature (Naessens et al. [2011]) as well as practice guidelines set by governments (MDH [2010]). However, we should note that other categorization rules can also be used if deemed appropriate.

To estimate the daily demand rate of a patient from each category, we calculate the probability that a patient from a certain category will request an appointment on a given day (λ_i^0 values), i.e., the total visits over a year for that category divided by the total patients in the category times the total workdays in a year. The daily appointment request rate for a given category is simply the multiplication of this

probability with the patient counts in that category. The total daily appointment request rate from a physician's panel is the sum of daily appointment request rates from all categories.

To estimate the physician service rate, we use the idea of adjusted service times based on case-mix mentioned above. Since no time-stamps data are available for us to estimate the length of provider consultation time, we set the following length based on the experiences of PCIM physicians. These lengths also seem to be consistent with those reported in the literature (Mechanic et al. [2001]). Patients with zero, one and two comorbidity count category require a 20 minute visit on average, whereas those that belong to higher comorbidity count categories require a 40 minute visit on average. Thus, we calculate the appointment duration for a physician by taking the weighted average of the service times. That is, we multiply the proportion of the patients that belong to the zero, one and two comorbidity count category with the required average appointment time (20 minutes) and add with the product of the proportion of those with higher comorbidity counts and the 40 minutes average. For example, if a physician has 50% of patients with lower comorbidity counts and 50% with higher, then the average appointment duration for this physician is $0.5 * 20 + 0.5 * 40 = 30$ minutes. Assuming eight hour work time every day, this physician can see on average 16 patients (=8 hours/30 minutes) daily.

3.2.3 Model analysis

Under our model assumptions, the first practice design becomes a simple M/M/1 queue and the third design is an M/M/2 queue; see Appendix A for the notation

and analysis of such models (Gross and Harris [1985]). The second design, however, is difficult to analyze. It is not amenable to the standard balance equation approach (Kulkarni [1995]), due to the inclusion of flexible patients. Gurumurthi and Benjaafar [2004] is the first to provide exact methods for the analysis of queuing systems with general customer and server flexibility and heterogeneous servers. Their analytical model allows asymmetric demand and service times, as well as an arbitrary flexibility matrix. The models they generate can be used to analyze flexible queuing systems in a variety of applications. Recently, Guo and Hassin [2012] study a two-server queuing system where some customers may place duplicate orders at both servers but will immediately withdraw one when they receive services from the other. More importantly, they are the first to provide closed-form formulas to analyze such a system. A close examination of their work reveals that their model is equivalent to our second practice design where the flexible patients play the role of customers placing duplicate orders in the queueing system. Thus we can adopt their formulas to analyze our second design. The formulas and the steps of the calculation for the waiting times is explained in detail in Appendix A. We use Microsoft Excel for the computations.

One of the primary benefits of using a queuing model is that it produces useful steady-state outputs. In our paper we only make use of some of them: utilization of the physician, probability that a patient will be seen by her own provider (continuity of care measure) and average waiting time for the patients (access to care measure). All these measures can be calculated using closed-form formulas reported in the literature discussed above.

3.3 Results

3.3.1 Impact of case-mix on provider utilization

System utilization is an important measure for the workload placed on a service system; it is evaluated as the ratio of patient daily request rate and the provider daily service rate. A higher utilization level indicates a heavier workload, i.e., more percentage of time being spent by providers in seeing patients; however, it also comes with more congestion and longer patient wait. More importantly, as the system utilization increases, the customer wait does not increase linearly but rather exponentially (Green [2011]). That is, when the system utilization is high, even a small disturbance, such as a slight increase in patient demand or drop in service rate, can significantly increase patient wait. Therefore, the system utilization is a crucial measure to monitor and control for, in order to balance the utilization and congestion in a service system. In this section we discuss how case-mix can affect this important measure.

To illustrate, we use Mayo comorbidity count visit rates, λ_i^0 values (see Table 3.1), to create seven hypothetical panels with the same system utilization 93.5% in Table 3.2. Recall that patients with different acuity may have different appointment demand rates, and they may also require different length of service times. Thus, it is not too surprising to observe that although these panels have the same system utilizations, their sizes are dramatically different due to different case-mixes. The largest panel is panel 5, in which a majority of the patients have no more than 4 comorbidities; in contrast, panel 4 is the smallest panel whose size is even smaller

than a quarter of panel 5, and it is predominantly occupied by patients with more than 4 comorbidities.

Table 3.1: Arrival rate per patient per day for each category.

Comorbidity count							
0	**1**	**2**	**3**	**4**	**5**	**6**	**7**
0.006	0.011	0.015	0.02	0.026	0.03	0.038	0.041

Table 3.2: Example of 7 hypothetical panels, with varying case-mixes, panel sizes, daily request rates and service rates. All the 7 panels have the same utilization of 93.5% (Rates are daily)

	Comorbidity count								**Panel Size**	**Arrival Rate**	**Service Rate**
Panels	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**			
1	160	150	226	200	142	42	14	3	937	15.53	16.6
2	100	100	226	161	108	40	50	20	805	14.98	16.05
3	50	50	140	140	5	30	85	90	590	13.67	14.57
4	5	5	5	5	25	50	100	125	320	11.34	12.17
5	425	350	275	200	110	13	2	1	1376	17.8	19.05
6	5	5	180	150	100	30	64	45	579	13.53	14.45
7	300	275	250	184	108	42	14	3	1176	16.89	18.05

As mentioned above, one way to balance workload and improve practice is via panel redesign. That is, reassigning patients across panels in the long term to achieve identical workload proportions and thereby using the existing capacity in the most efficient way possible (Balasubramanian et al. [2010]). Here we use two real physician panels from Mayo Clinic Primary Care Internal Medicine (PCIM) to demonstrate the effect. The initial panel size and case-mixes are shown in Table 3.3. Physicians 1 and 2 differ in their case-mixes, panel sizes, arrival and service rates and therefore utilizations. Physician 1 has a utilization of 94.8%, while Physician 2 has a utilization of 99.6%. These differences can occur in practice due to reasons such as physician

seniority, physician and patient preferences. As a result, patients of Physician 2 will experience poorer access compared to those of Physician 1. Now, what if the panels could be redesigned such that these two physicians had similar case-mixes? In this case, we balance panels simply by dividing the patients from each comorbidity count category equally among the two physicians. In doing so, the utilization of each physician equals at 97.2% (see Table 3.4). In the next section, we will discuss how panel redesign affects the access to care and continuity of care measures.

Table 3.3: Case-mixes of Physicians 1 and 2: Initial/Baseline panels (where PS: Panel Size,RR: Request Rate, SR: Service Rate)

	Comorbidity count											
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**PS**	**RR**	**SR**	**Utilization**
Physician 1	380	372	269	187	98	33	8	1	1348	17.91	18.91	94.70%
Physician 2	230	272	240	190	124	47	23	5	1131	17.38	17.45	99.60%
Total	610	644	509	377	222	80	31	6	2479			

Table 3.4: Case-mixes of Physicians 1 and 2: Balanced Panels/After redesign (where PS: Panel Size, RR: Request Rate, SR: Service Rate)

	Comorbidity count											
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**PS**	**RR**	**SR**	**Utilization**
Balanced	305	322	255	189	111	40	16	3	1241	17.68	18.18	97.20%

3.3.2 Comparison of practice designs under different case-mixes

In this section, we compare the three practice designs introduced before (see Figure 3.1). Recall that in Design 1, the two physicians practice independently; while in Design 3, they form a provider team and share all their patients. In the former case we expect to see long waiting times (i.e., poor access) especially for a highly

utilized physician, but the continuity of care is perfect for all patients. However, in the latter case, we expect the waiting times to decrease but continuity of care is no longer perfect. The patients may see one of two providers and hence continuity is 0.5 as opposed to 1 in the first case (0.5 means that patients will be seen by their own providers with 50% chance).

Between these two extremes of best continuity and best access is the partial pooling case, i.e., Design 2, where the providers form a team and share a subgroup of patients. Care for this group of patients could be provided by either provider; continuity for the shared patients is therefore 0.5. But each provider also retains a certain number of dedicated patients for whom continuity is 1. Thus, based on the number of patients shared and the number of patients dedicated, we can calculate an overall (weighted) continuity of care measure. If 50% of the total visits are shared by the two providers, and 25% are dedicated with each of the physicians, then the weighted continuity measure is $1 * 0.25 + 1 * 0.25 + 0.5 * 0.5 = 0.75$.

In practice, it makes sense to provide greater levels of continuity to patients with multiple chronic conditions. Reid, R. J. and Coleman, K. and Johnson, E. A. and Fishman, P.A. and Hsu, C. and Soman, M.P. and Trescott, C. E. and Erikson, M. and Larson, E.B. [2010] and Coleman et al. [2010] discuss that in Group Health Practice during the reassignment of panels, when physicians were given the chance to choose which patients to keep, they preferred the elderly and sicker patients. Compared to relatively healthy patients, these patients need a stronger bond with their PCP for better management of their health conditions.

To start with, we allow the providers to share only patients with zero comorbidity count, i.e., CC = 0, who are apparently healthier patients in the panel; all other patients still remain dedicated to their respective providers. We calculate access and continuity measures for this setting. Next we allow providers to share patients with CC up to 1, thereby increasing the number of shared patients and again calculate access and continuity measures. We proceed in the same way until all patients are shared by the two providers; this becomes Design 3. In our data, since CC range from 0 to 7, we have a total of 9 cases, including the two extreme cases (i.e., Designs 1 and 3).

Table 3.5 provides the waiting time and continuity measures for each of these 9 cases, for baseline panels and panels balanced via redesign introduced in the last section. Figure 3.2 summarizes the changes in access and continuity of care provided across all 9 cases for both the baseline and balanced panels. In Table 3.5, W1 is the average appointment delay of patients dedicated to Provider 1; W2 is the average appointment delay of patients shared by both providers; W3 is the average appointment delay of patients dedicated to Provider 2. Clearly in the M/M/1 case, since no patients are shared, W2 does not exist. Similarly, since no patients are dedicated in the M/M/2 case, W1 and W3 do not exist. W is a consolidated access measure for all patients, calculated as the weighted average of W1, W2 and W3, where the weights are based on the proportion of the arrival rates for the dedicated and shared patients. The unit of W1, W2 and W3 is days.

In the baseline dedicated case (Design 1), Provider 2's patients have average appointment delay of 13.8 days (see W3), while Provider 1's patients have an average

Table 3.5: Design comparison under the baseline and balanced panels, where CG: Comorbidity groups, WC: Weighted Continuity

CG shared	% pooled	WC	Baseline Panels				Balanced Panels			
			W1	**W2**	**W3**	**W**	**W1**	**W2**	**W3**	**W**
None (Dedicated)	0%	1	1	0	13.8	**7.3**	1.8	0	1.8	**1.8**
0	11%	0.95	1	0.9	1.2	**1.1**	1.1	0.9	1.1	**1**
0-1	30%	0.85	0.9	0.8	1	**0.9**	1.0	0.9	1	**0.9**
0-2	52%	0.74	0.9	0.8	0.9	**0.9**	0.9	0.9	0.9	**0.9**
0-3	73%	0.64	0.9	0.8	0.9	**0.9**	0.9	0.9	0.9	**0.9**
0-4	89%	0.55	0.9	0.9	0.9	**0.9**	0.9	0.9	0.9	**0.9**
0-5	96%	0.52	0.9	0.9	0.9	**0.9**	0.9	0.9	0.9	**0.9**
0-6	99%	0.5	0.9	0.9	0.9	**0.9**	0.9	0.9	0.9	**0.9**
0-7 (Pooled)	100%	0.5	0	0.9	0	**0.9**	0	0.9	0	**0.9**

delay of only 1.0 days (see W1). This dramatic difference is due to the imbalance in the case-mixes of these two physicians, which results in 99.6% utilization for Provider 2 and 94.8% utilization for Provider 1, as discussed in the last section. This also signifies our earlier point that when utilization level is high, a slight increase in utilization will lead to a dramatic increase in patient wait. Now, if we look at all patients, the average delay is 7.3 days in this case and the continuity of care is perfect. However, if we were able to redesign the panels of these two physicians and balance their workload, the utilization of both providers equals at 97.2% and the average appointment delay for all patients is reduced to 1.8 days (see Balanced Dedicated case). This is a 75% improvement in access to care.

Next, consider the Baseline panels when 0 Comorbidity Count (CC) patients or apparently healthy patients are shared by the 2-physician team. The access improves significantly for all patients (see W1, W2 and W3), with the overall average delay reduced from 7.3 days to 1.1 days (85% reduction). Interestingly, the overall continu-

ity measure drops only marginally from 1 to 0.95. Thus for a 5% drop in continuity of care in relatively healthy patients, we get an 85% improvement in access to care. For the balanced panels after panel redesign, we obtain similar findings. On top of the benefits generated from panel redesign, pooling 0 CC patients further reduces overall patient waiting time by additional 44% (from 1.8 days to 1.0 days) with only 5% drop in continuity measure.

A closer examination of Figure 3.2 reveals that when more patients are shared by the two physicians, access measures improve, but the improvement is not as significant as going from the dedicated to the 0 CC shared case. Furthermore, as more patients are shared, the Baseline and Balanced cases tend to get similar. When all patients are shared, they converge to Design 3 and have the same access and continuity measures.

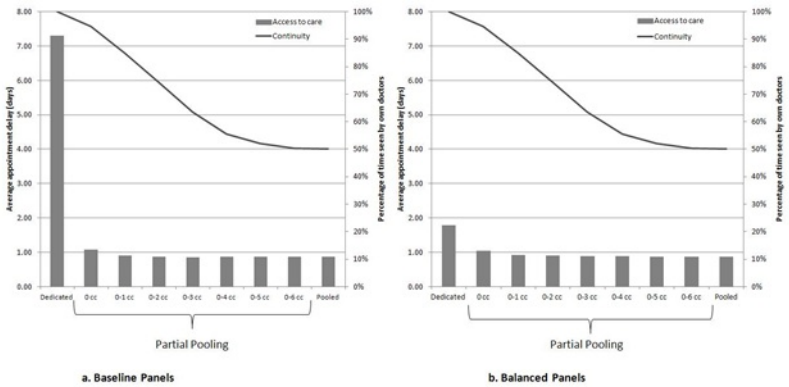


Figure 3.2: The impact of partial pooling on access to care and continuity of care for both baseline and balanced panels. The x-axis ranges from the fully dedicated case (Design 1) to fully-pooled case (Design 3)

3.4 Discussion

Our study is among the first to develop case-mix adjusted methods to evaluate continuity of care and access to care for primary care delivery models. We consider three commonly-used practice designs, namely dedicated service design, partial pooling design and complete pooling design. Our study highlights the importance of considering case-mix in primary care practice design. Case-mix not only affects, on average, how frequently patients need healthcare services, but also influences how much time/ resources that a patient needs for each visit. Many primary care providers in the U.S. have panel sizes exceeding 2000 patients regardless of the case-mix (Alexander et al. [2005]), Green et al. [2013] even look at alternative methods of delivering care on the supply side in order to increase the nationwide panel size of 2500 patients to 5000 patients due to the soon to increase demand in primary care. Our results suggest that such a seemingly one-size-fit-all approach does not work. Providers can easily feel overwhelmed if their panels contain a relatively large number of patients with complicated conditions. It is crucial to take case-mix into account.

A practice typically has two strategies to improve access to care with available capacity. One is to create provider teams and pool service capacity on a certain group of patients. This strategy seems to be spreading fast in the U.S. as more and more primary care practices shift from solo-practice to group practice and physicians cover each others work in a care team. Indeed, the share of solo practices fell to 18 percent by 2008 from 44 percent in 1986, according to the AAFP's 2008 member survey (Harris [2011]). One important question that arises from such a practice shift is what kind of patients can be shared. Intuitively, patients who are relatively

healthy can be shared because their cases are relatively simple and easy to handle. In our data examples, we use comorbidity counts (CC) as a measure for patient health status and the provider can choose who to share based on their CCs. We find that, letting providers share patients with 0 CC, who only contribute to 11% of the total visits, can significantly improve access to care by 85% but continuity of care only decreases by 5%. More importantly, most of the benefits that can be generated by patient sharing come from just sharing zero comorbidities patients. In other words, a little flexibility can go a long way. Indeed, such ideas of using flexibility have been discussed in other non-healthcare contexts such as manufacturing (Jordan and Graves [1995]), and are shown to be effective in improving system efficiency.

The other strategy often used by practice is via panel redesign to balance workload among physicians. In our data examples, the two physicians have imbalanced panels at baseline. Panel redesign alone can improve the overall access of care by 75%. However, when a practice tries to redesign existing panels, it usually involves much effort related to redirecting and re-empanelling the patients; and such changes can take a long time and much effort (Balasubramanian et al. [2010]). The reassignment experience at the Group Health practice in Seattle also illustrates these challenges (Coleman et al. [2010]). Instead, if panels were to be designed proactively in the early phase of empanelling new patients rather than to be redesigned reactively after panels have been formed, the work might have been much easier and effective.

There are other strategies that a primary care practice can use to improve access to care. For instance, some practices choose to delegate certain tasks, e.g., preventive care and chronic care work, to non-physician members of the care team. A recent

study examines how such task delegation affects the choice of panel size (Altschuler et al. [2012]). In particular, it considers how much time a physician can save by task delegation, and then simply equates the available physician time with the time consumed by patients to derive the reasonable panel size, which usually gets expanded post task delegation. Altschuler et al. [2012] do not, however, consider the impact of such system changes on continuity or access to care. In contrast, our modeling framework can achieve both ends, i.e., considering task delegation and evaluating continuity and access to care. To do so, we just need to include only patient visits to the physician in our model analysis.

Our modeling framework is developed using queuing theory. It provides general and yet easy-to-use tools to model and analyze service systems when customer wait is an important focus of the problem. Despite its many merits, this method also has a few limitations in modeling a primary care practice. In particular, we assume that the service process is continuously running and "ignore" weekends when most practices are closed. We also assume that patients are always assigned to the earliest appointment slot available although it may not be the case in reality. Thus the appointment delay estimates generated by queueing models may underestimate the actual patient wait time. Using the weighted average of the service times is an approximation but this allows us to use tractable, closed-form expressions. As an extension, for a more accurate estimation Gurumurthi and Benjaafar [2004]'s novel approach can be implemented.

However, our analysis depicts how the appointment delay varies across different practice designs (see Figure 3.2), thereby enabling us to evaluate the relative changes

in access to care. These relative changes perhaps provide more useful information compared to the absolute values of appointment delays when comparing practice designs.

Our study points to several future research directions. First, we use comorbidity count as a criterion for patient sharing. It is important for the clinical community to study how patient sharing affects health outcomes and develop guidelines for it, i.e., who to share or when to share. Second, it will be interesting to develop simulation models (Law and Kelton [1991]) rather than analytic models (like ours) to study different primary care practice designs. The advantage of a simulation model is that it can incorporate more details and represent the reality better; however, it is usually developed based on a single facility, making its results difficult to generalize. Third, our models only consider primary care providers, e.g., physicians and nurse practitioners. There are many other important medical professionals in a care team, e.g., medical assistants. It will be interesting to develop more comprehensive models to study the dynamics and patient flow through the whole care team. Last but not least, we only consider the effect of panel redesign. How to proactively develop panels in the early phase of building up a group practice remains an unexplored and yet very important research topic.

CHAPTER 4

MODELING HOSPITAL-WIDE PATIENT FLOWS USING SIMULATION

4.1 Introduction

Hospital care accounts for 31% of the nation's health expenditures (Martin et al. [2012]) and inpatient beds are one of the most important resources in a hospital. A mismatch between demand and supply in inpatient beds can cause hospital wide congestions. Green [2003] and Williams [2006] point that the unavailability of inpatient beds affects the functioning of other parts in the hospital. These effects include but are not limited to: patients waiting long hours in the emergency department (ED) for an inpatient bed; patients not being placed in their primary unit (i.e. off-service placement); urgent patients bumping less critically sick patients from intensive care units (ICUs) to "step-down units"; patients waiting in post acute care unit (PACU) for an inpatient bed, operating room (OR) delays; ambulance diversions and refusing transfers from other hospitals.

In this paper, we use an empirically calibrated discrete event simulation to quantify the impact of discharge timing on timely access to inpatient beds. By discharge timing we mean how the number of patient discharges varies by the hour of the day. Timely access is measured in two ways: 1) by the average non-value added waiting time spent by a patient in the ED, PACU or other locations after the physician has made a request for an inpatient bed; and 2) the average number of patients waiting for an inpatient bed (average queue length).

Our simulation model is based on a year's worth of inpatient flow data from Baystate Medical Center (BMC), an acute care medical center in the Northeast of the U.S. On average each day there are around 100 bed requests and discharges at this medical center. Figure 4.1 shows the mean number of inpatient bed requests and patient discharges by hour of the day at BMC. The time-varying nature of the admission requests and the discharge process can be clearly observed. Notice that discharges peak in the afternoon between 2-4 PM, producing a bell-curve centered on these afternoon hours.

There are 2 main reasons behind the underlying empirical discharge distribution as observed in Figure 4.1, which are hospitalist shifts and prioritization rules. First, the hospitalist shifts in the hospital is divided into 3 shifts. During regular hours (8 AM to 5 PM) when most of the discharges happen, around 15 to 18 hospitalists are scheduled; after 5 PM to midnight there are only 2 hospitalists and very few discharges; and after midnight, there is typically just one hospitalist for urgent cases and no discharges happen in this duration. Second, during 8 AM to 5 PM, when hospitalists are doing their rounds they tend to first see the recently admitted patients

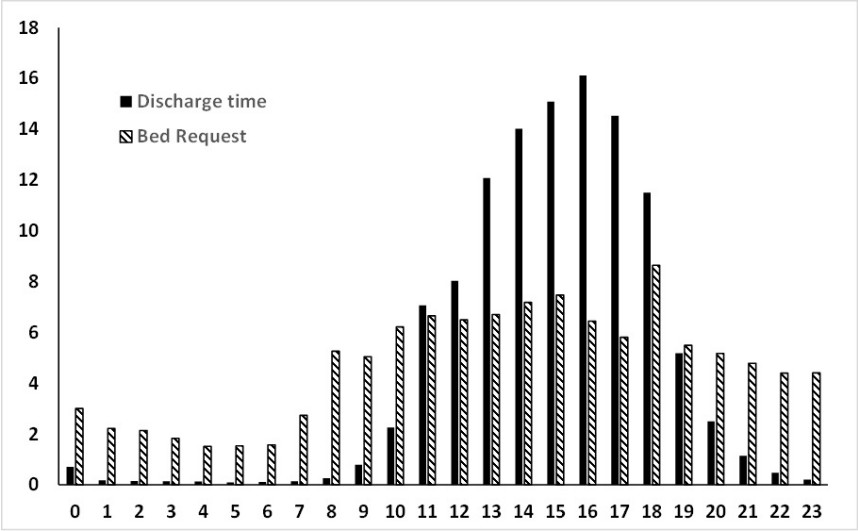


Figure 4.1: Admission and discharge rates

or patients who are critical, which results in discharges that are completed later in the day with a higher LOS.

Bed requests also vary by the hour, although less dramatically than the discharge rates. Once a bed request is fulfilled, a patient stays on average for 4.8 days (around 120 hours). Thus patients who are being discharged, say today, most likely requested for an inpatient bed a few days ago. This two time-scale feature distinguishes the inpatient admission and discharge process from the service settings typically studied in the operations research literature. We return to this point again while reviewing the relevant literature.

We investigate in this paper whether discharge profiles different from empirically observed one in Figure 4.1 (our baseline) can improve timely access to inpatient beds. A discharge profile is defined by (a) discharge window, which specifies the hours of the day discharges are allowed; and (b) maximum capacity for discharges in each

hour of the window. The typical window is 10 AM-7 PM, and capacity varies in each hour of the window as shown in Figure 4.1. A number of alternatives are possible to ensure speedier admissions for waiting patients. For instance, how much would be gained if the hospital tried to discharge most of its patients by noon, thereby freeing up beds earlier in the day? Are discharges by noon feasible given current capacity constraints of the hospital? What if a more uniform discharge capacity was adopted from one hour to the next or if discharge hours were extended in the evening by a few hours? How much would waiting be impacted if the hospital tried to prioritize discharges in units that had longer queues? We also discuss the feasibility of these alternatives in practice for both patients and hospital staff. Some of the issues are discussed in Tables 4.1 and 4.2.

We consider heterogeneity in inpatient bed request sources (such as ED, surgical area, community referrals), clinical diagnostic categories, and desired inpatient unit (medical-telemetry, renal, psychiatric unit and so on). Inpatient length of stay (LOS) of the patients in our model varies as a function of these categories. We also consider time-varying (non-homogeneous) inpatient bed request rates. With these as inputs, we use the model to test a wide variety of discharge profiles in our model.

Together, this constitutes a time-varying multi-server queuing network model with multiple patient classes. The model is queuing network for two reasons. First we allow patients to first visit the ICU before stepping down to regular unit. Second, in our model there is a front-end queue of patients waiting to get admitted to an inpatient bed, and a back-end queue of patients who have completed their LOS and are now waiting to be discharged. The front-end queue builds up in each unit

Table 4.1: Earlier in the day discharge

	Pros	**Cons**
Patient's Perspective	Go home earlier in the day.	Might be thought of a premature decision. Timing might be hard for families, since they will have to leave work to pick up the patients during the day.
Hospital's perspective	Patient flow improves significantly as the beds free up before the demand for it builds up. Better financially, since the earlier the patients leave the hospital, they need to provide less food and medicine.	A big burden on the hospital and hospitalists to coordinate most of the discharges to happen before noon. Unrealistic.
Physician's perspective	Patient will be home early in the day if there are any questions, issues– and MD office would still be open– avoid night time questions to MD's office.	The physician/ hospitalist would need to address discharge issues (expected to be routine matters). This competes with the need to see new admits, sicker patients, and patients with issues occurring during the night.

since beds may not be available, and the back-end queue may develop since the discharge capacity of the hospital (hospitalists in our case) in a particular hour may be tight. What's more, this capacity is expected to grow tighter since the demand for hospitalists nationwide is expected to grow, as life expectancy is increasing and

Table 4.2: Later in the day discharge

	Pros	**Cons**
Patient's Perspective	Family members might be more available to pick up the patient later in the day than during the day when they have to leave work. Also patients will go home rather than having to stay another night in the hospital.	Leaving the hospital at night might be inconvenient for some patients, who need the delivery of oxygen or medical equipment. Some pharmacies close early and prescription pick up could be complicated in the evening.
Hospital's perspective	Easier to coordinate than early in the day discharge.	Hospitalists shifts and hours of the ancillary services will need to readjusted.
Physician's perspective	Allow time to have all test and lab results from day. Hospitalist can speak directly to family members who work during the day.	Any problems that arise when the patient gets home, it is "after hours" to reach physician. Could be managed by hospitalist being available for any follow up calls from patient.

older, sicker patients mean more complex case management for hospitalists (Collins [2012a]).

This two service line feature gives us the opportunity to test whether prioritizing discharges for those units that have the longest admission (front-end) queues has an impact on timely access to inpatient beds. Prioritization thus allows us to model state-dependent discharges, where a hospital responds in a holistic fashion by recommending that physicians and support staff (nurses, case managers) conduct their rounds and other discharge related preparations in units which have more patients

waiting to be admitted. We reiterate that we are not recommending a patient be discharged before their LOS is completed. Discharges in our model only apply for patients who have completed their LOS.

We quantify the impact on admissions queue lengths for each unit under various discharge profiles with and without prioritization, and thereby study the individual and combined effects of these two factors. Even though, early in the day discharge policy has been studied extensively in the literature (Shi, P. and Chou, M. C. and Dai,J.G. and Ding, D. and Sim, J. [2012], Powell et al. [2012]), to the best of our knowledge, we are the first to propose and evaluate the impact of extending regular discharge hour windows. Early in the day discharge policy alone does not result in significant improvement in waiting times, and it also requires significant behavioral change. However, expanding the discharge windows by only 2 hours creates the same benefit with early in the day discharge policy. We also model a more responsive discharge policy that prioritizes units in allocating the restricted discharge capacity based on the admissions queue. This prioritization scheme results in significant improvements in decreasing waiting times.

A less tangible but equally important contribution is the fact that the entire simulation modeling process - assumptions, data inputs, analysis of outputs, implications for practice, implementation of results – was conducted over a 3-year period with constant input provided by key stakeholder groups at our partner hospital.

This chapter is organized as follows. We first provide some background on discharge planning and then review the relevant literature, and make the case for why we did not choose queuing models to tackle the problem (Section 4.3). We then

describe the data and provide basic background on BMC's operations in Section 4.4. After building a simulation model that mimics the patient flow in our partner hospital in Section 4.5, we evaluate different discharge profiles both quantitatively and qualitatively (Section 4.7). In Section 4.9 we discuss some potential future directions. As an ongoing work, in Section 4.10 we present our motivation for modeling overflow transfers with some preliminary analysis. Lastly we discuss a related future research direction: hospitalist scheduling problem, in Section 4.11.

4.2 Discharge Planning

Medicare describes discharge as "a process used to decide what a patient needs for a smooth transition from one level of care to another". In general, the basics of a discharge plan are: (1) Evaluation of the patient by qualified personnel, (2) Discussion with the patient or her representative (which includes details of the types of care that will be needed; and whether discharge will be to a facility or home; information on medications and diet; what extra equipment might be needed, such as a wheelchair, oxygen tank and so on); (3) Planning for homecoming or transfer to another care facility; (4) Determining if caregiver training or other support is needed; (5) Referrals to home health agencies and/or appropriate support organizations delivering needed equipment to the home; and (6) Arranging the follow-up appointments or tests (FCA [2013]).

The timing of discharges is closely related to how hospitalists prioritize patients on their rounds. The term "hospitalists" was first used in 1996 (Wachter and Goldman [1996]). Hospitalists are specialists in inpatient medicine, and are responsible for

managing the care of inpatients in the same way that PCPs care for outpatients. They admit the patients to the hospital, plan their workup, and arrange the transition back to the outpatient setting (Wellikson [2010]; Maguire [2009]).

In the morning rounds, hospitalists need to prioritize their work and although they can discharge patients who are ready, such a policy violates their first rule of triage: "see the sickest (the newly admitted patients) patients first". Priority overall is given to admissions and acute patients (Quinn [2011]), not to discharges. This results in discharges either being deferred or completed later with a greater LOS. Patients who are ready to go home, although relatively less sick, need the hospitalists' attention as well to start the discharge process which involves initiating paperwork, ordering tests, educating the patient and developing a care plan for discharge.

Late discharges are typically the result of the timing of physician rounds, lack of coordination with the patients' family members about the discharge time and delays resulting from post-acute care facilities. University of Utah Hospitals and Clinics have identified thirty non-medical barriers to a timely discharge, with transportation (28%), late discharge order (13%) and patient delay (8%) being the three major reasons (Nelhin [2006]). The patients to be discharged on a given day are typically known the day before. But even if these patients are ready to leave on the morning, their discharges happen much later in the day.

Besides having adverse effects on patient flows, delayed discharges have clinical drawbacks like increasing the possibility of hospital acquired infections (DH [2004]). Also the hospital has to provide nursing care, food and medicine until patients are discharged, so it creates a further financial burden on the hospital. What's more

patient satisfaction is affected adversely. Thus, discharging medically fit patients in a timely manner has many potential benefits and can improve the waiting times of patients for a bed significantly.

Although there are only a few papers that have looked at the relationship between discharge timings and waiting times (Shi, P. and Chou, M. C. and Dai,J.G. and Ding, D. and Sim, J. [2012], Powell et al. [2012], Bekker et al. [2014]), a possible solution to inpatient congestions is implementing effective discharge policies.

One of the main discharge policies of interest is the early in the day (EITD) discharge policy, as it has both been proposed in the literature and applied in practice. For example, Hospital of Miami has set a goal of discharge time by 11 AM to free beds earlier in the day (HMA [2006]), by having specific nurses who work as patient discharge care facilitators. Their main job is to do rounds with hospitalists to identify patients who will be ready to be discharged the next day and get a running start on the discharge to-do list (tests, paperwork, patient educations, scheduling follow-up appointments and arranging transportation) in order to discharge patients earlier in the day. They also point out that no matter how early the team starts to work there are generally delays when transporting patients to nursing homes.

Research in this field presents conflicting outcomes. Shi, P. and Chou, M. C. and Dai,J.G. and Ding, D. and Sim, J. [2012] demonstrate that this policy does not significantly improve the hospital operations. On the other hand, the results from Powell et al. [2012] show the opposite. Powell et al. [2012] test the impact of shifting the discharge distribution to earlier in the day as well as the impact of two inpatient discharge timing policies, by using a simple Excel model, with homogeneous demand

and total bed flexibility. They conclude that the alternative discharge policies decrease both the ED and surgical boarding (waiting) time. One of the most drastic results is that even by shifting the peak discharge hour from 3 PM to 2 PM there is a decrease in waiting time by 50%. However, an important caveat is that their model is static and deterministic, whereas Shi, P. and Chou, M. C. and Dai,J.G. and Ding, D. and Sim, J. [2012] report on an actual implementation (at a Singapore hospital) of an early discharge policy.

4.3 Literature Review

The literature is reviewed in two parts: first we go over the literature for hospital-wide flow models and secondly discuss why we used a simulation model as opposed to a queuing model by reviewing a list of queueing models applied to healthcare networks with time-varying arrivals.

4.3.1 Hospital-wide flow models

Modeling and improving patient flow has been studied extensively in the literature. In fact, the problem of inpatient bed congestion is not only prevalent in the U.S., but it is a problem commonly observed in other countries like Singapore and Israel (Armony et al. [2012], Shi, P. and Chou, M. C. and Dai,J.G. and Ding, D. and Sim, J. [2012]). Much work has been done in this area (see Williams [2006] for a detailed literature survey), but we are specifically focusing on the hospital-wide optimization models rather than unit specific ones.

For instance, a common way to address the problem is by implementing better OR schedules in order to reduce census variability. The motivation is that elective admissions typically exhibit more census variability than ED patients. This is due to a lack of consideration of downstream effects, ward/bed requirements in admit decisions (Helm and Van Oyen [2010]). As a solution Bekker and Koeleman [2010] use a combination of quadratic programming and queuing theory in order to come up with quota scheduling for elective surgeries to reduce the artificial variability caused by scheduled surgery patients. Whereas, Helm and Van Oyen [2010] look at this problem by using a "Poisson-arrival-location" model (PALM) based on patients' stochastic location, and further develop a deterministic model using probability distributions for patient pathways. They find the optimal mix of elective patients that will smooth the census by coordinating with the admit decisions in the hospital.

Even though this approach smoothes the bed census, the interactions between different demand lines are ignored. Unit specific analysis might optimize a specific part of the hospital but will not consider impact on the hospital as a whole. Thus, we turn our focus to hospital-wide optimization models. Various IEOR techniques have been used including queuing models (Bekker and Koeleman [2010], Shi, P. and Chou, M. C. and Dai,J.G. and Ding, D. and Sim, J. [2012], Armony et al. [2012], Green and Nyugen [2001], Green [2003]), mixed integer programming (Helm and Van Oyen [2010]), Markov decision processes (Helm et al. [2011], Helm et al. [2010]), stochastic optimizations (Best et al. [2012]), in order to alleviate the inpatient bed congestions.

However, by far the most common methodology used for modeling and improving inpatient flow in hospitals is discrete event simulation, since Hancock and Walter [1979]. Hancock and Walter [1979] have used simulation for inpatient admission scheduling in order to reach the maximum occupancy attainable. Since then many papers have used discrete event simulation as a decision support system in order to simulate flows in a hospital (Montgomery and Davis [2013], Proudlove et al. [2007], Helm et al. [2011], Shi, P. and Chou, M. C. and Dai,J.G. and Ding, D. and Sim, J. [2012]). For example, Proudlove et al. [2007] have used a forecasting and simulation model to gain generalized insights and to be able to demonstrate "what if?" scenarios rather than reproducing a specific small part of the network. Helm et al. [2011], on the other hand, develop methods to control the inpatient admissions to decrease the negative impacts of demand variability, by using a combination of Markov Decision Process (MDP) and a simulation model. Also the papers using simulation models, almost always highlight that cooperation and careful planning with stakeholders, and simulation model graphics or visualization are crucial aspects for implementation aspect (Forsberg et al. [2011]).

4.3.2 Why simulation and not queuing?

Both queueing and discrete event simulation models have been used extensively in improving and modeling healthcare problems. Queuing models and simulation models each have their benefits. Queuing models are simpler, require less data, and provide more generic results than simulation (Green [2006]). On the other hand, dis-

crete-event simulation is more flexible and enables us to model the details of complex patient flows.

Kolker, A. [2010] provides examples that clearly demonstrate why in most cases discrete event simulation models are superior and preferred over queuing models. Kolker looks at an example in a healthcare setting with time-varying arrival rates, and concludes that the queuing analysis should not be used for such models. Supporting this, Green [2011] discusses that for these types of queuing systems "using queuing models is inappropriate for estimating the magnitude and timing of delays, and a simulation model will be far more accurate".

The staffing problem lends itself easily to queuing models, because of its closed form expressions for useful output measures like waiting times, number of people in the queue. This is one of the main reasons why queuing models have been used extensively in healthcare as well. For example, Green [2003] evaluates the optimal bed capacity based on a target probability of delay using an M/M/S queuing model. Green and Nyugen [2001] use a queuing model to determine optimal policies for bed planning considering the trade-off between delays and occupancy levels. Though simple models, they are tractable and develop insights on hospital capacity planning that are generalizable.

Despite the abundant literature on stationary queuing models applied to healthcare processes, research on applications of non-stationary arrival rates is scarce. Closed form expressions typically do not exist for non-stationary customer arrival rates. Numerical analysis, stationary model approximations, infinite server approximations and fluid approximations have typically been used to generate approxi-

mations. To the best of our knowledge, the only papers that consider time-varying arrival rates in modeling healthcare queuing networks are Armony et al. [2012], Green et al. [2007b], Yom-Tov and Mandelbaum [2010] and Zeltyn et al. [2009]. The closest formulation is Shi, P. and Chou, M. C. and Dai,J.G. and Ding, D. and Sim, J. [2012] who develop a two-time scale queuing network but they are only able to model one patient source rather than multiple patient sources, a feature essential to our hospital flow model.

One of the most extensive queueing research in inpatient flows is conducted by Armony et al. [2012], who analyze the hospital-wide patient flow from a queuing approach. Exploratory data analysis (EDA) is used to study detailed patient flow data from a large Israeli hospital. They analyze the flow in ED, internal wards (IWs) and the transfers from ED to the IWs. They emphasize the importance of understanding the system's behavior at hourly resolution. However, Armony et al. [2012] do not analyze the impact of discharge policies, but their main focus is on patient flow from ED to internal wards.

Shi, P. and Chou, M. C. and Dai,J.G. and Ding, D. and Sim, J. [2012] is the first to explore stochastic models to analyze the impact of effective discharge policies. To do so, they develop an analytical model and then use a very rigorous simulation model that mimics the inpatient operations in a Singaporean hospital. They study the effect of early in the day discharge, that was implemented in the hospital, on ED waiting. Similar to our finding in Section 4.7, the authors observe that this policy alone has limited impact on reducing waiting time. One of the major findings is that instead of an early in the day discharge policy, a hypothetical discharge distribution,

which still discharges 26% of patients before noon, but shifts the discharge peak time to 8-9AM, provides significant improvement in waiting times. Shi et al. also find further improvements by combining the impact of discharge policies with other policies like increasing the number of beds.

Motivated from their stochastic network, Dai and Shi (2014) develop an analytical framework with a two-timescale analysis. They evaluate time-dependent performance measures for a single class time-varying queuing network, while modeling the hospital inpatient flow. Using a stationary queuing system, they first obtain the performance measures on a daily level for the "midnight customer count". Whereas, the second time scale is used to derive the distribution for hourly customer count, which leads to the calculation of time-dependent performance measures for the single-customer class model (Shi, P. [2013]).

Our problem, on the other hand, consists of multiple patient categories each exhibiting a different LOS distribution and time-varying arrival process. We have looked at implementing time-varying arrival rate queuing models to our inpatient flow model. However, the queuing models are unable to tackle the complexity of the problem. As discussed in Shi, P. and Chou, M. C. and Dai,J.G. and Ding, D. and Sim, J. [2012], time-varying models have been studied extensively for call centers. Time between arrivals (TBA) and time between services (TBS) happen every few minutes in call centers – i.e. same time-scale – whereas TBA in inpatient setting is hourly and TBS is at least one day (on average 5 days). Clearly the latter case involves two different time-scales, as discussed in Ramakrishnan et al. [2005] and Shi, P. [2013].

Also, unlike call centers our model has extremely long service times and the number of servers (beds) cannot be adjusted in a short time window. Additionally, approximations for the time-varying queuing models, work well for under-loaded models, whereas typically at peak hours, the hospitals are working over capacity (using hall beds, unlicensed beds,...). Thus, existing approximation methods generated for call center models are not applicable to our hospital model. As a result, we have turned our focus to simulation models based on sampling from historical data collected from BMC, a large tertiary care hospital in the Northeast of the U.S. in our case.

The queuing framework behind the simulation model can be observed from Figure 4.2. There are time-varying arrival rates from different patient sources which can be categorized into 2 major sources as controllable (scheduled) and uncontrollable (urgent). The patients get admitted to units and require care depending on their MDC category, which are individual queuing systems (G/G/Cs) themselves. We only account for the ICU transfers, due to lack of data. After the LOS is complete, we assume the patients join the discharge queue. As can be observed we model two different service lines: one to be admitted and the other to be discharged. This allows us to model both a state dependent and an independent discharge profile and quantify the impact. Since, we have used a simulation model for modeling the inpatient flow process, we were able to construct a detailed system, as opposed to having numerous simplifying assumptions in analytical model formulation (Davies and Davies [1995]).

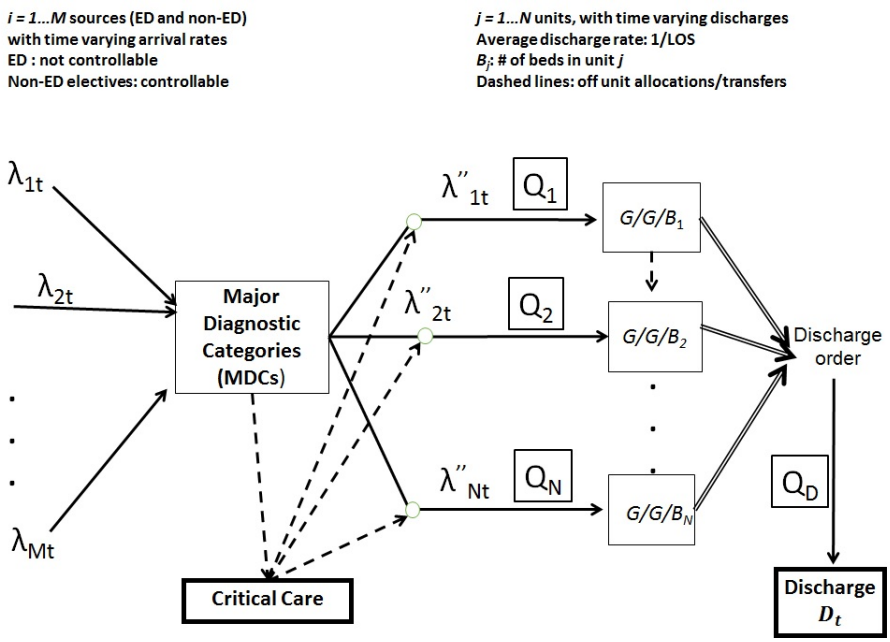


Figure 4.2: Queuing framework

4.4 Data Collection and Analysis

We analyzed data from all patients who were admitted to an inpatient bed at BMC from May 2010 to April 2011. We used anonymous patient records which included patient age and gender, and diagnoses related categorizations. These include the diagnosis related groups (DRGs) and major diagnostic categories (MDCs). This MDC categorization was initially created for the claims and administrative process; each MDC aggregates related DRGs into a single broader category – for example, two such categories are "Respiratory Diseases" and "Circulatory Diseases". There are 25 MDCs and this keeps the model concise and tractable. Additional data analysis for MDCs and the features of the data used in sampling is provided in Appendix B.

We have also analyzed the time-stamps for each patient and in fact these form the basis of some key inputs in our simulation model (see Figure 4.3). A patient may enter the hospital information system by registering through the ED, surgical unit, a physician's office or other sources. After the patient goes through the assessment, consultation and care process, the relevant physician or care provider decides that the patient should be admitted to an inpatient bed in a desired unit. This is the bed request time and in our simulation model it translates to a patient arrival.

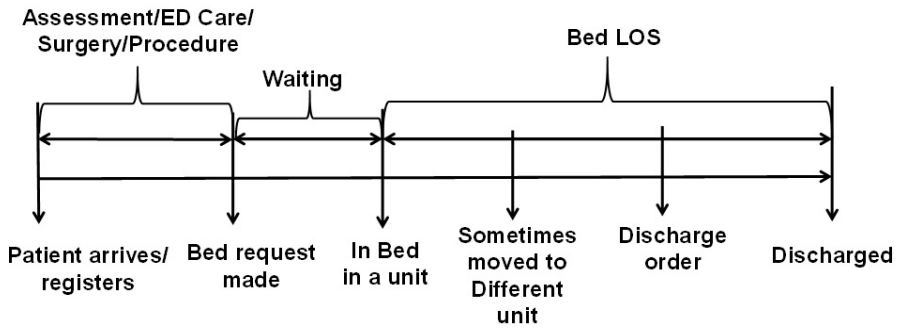


Figure 4.3: Admission and discharge process

The patient then waits until a bed is available and is then admitted. After staying for some duration in the inpatient bed, the patient is discharged. The important point here is that by length of stay we mean time spent by the patient in the inpatient bed. In Figure 4.3 this is "the discharge time" minus the "in the bed" time. From the point of view of our simulation model, inpatient bed LOS is the "service time" and number of inpatient beds in a unit are the number of "servers".

Unfortunately, we did not have data on patient transfers between units. Overflow transfers happen because the patient was originally admitted to a non-primary unit that may not have had the equipment and staff to adequately deal with the patient's

condition. Transfers in general are not desirable and are costly (West [2010b], West [2010a]). In fact, many hospitals are trying to implement a "right patient, right bed" policy for accommodating patients in the correct place, so they do not have to be transferred (West [2010a]). We assume in our model that 1) the unit from which the patient was discharged was the patient's desired unit; and 2) that there are no transfers other than critical care unit transfers: the patients simply wait in a non-ideal location until a bed becomes available in the desired unit. We have performed some preliminary analysis on overflow transfers, but will only focus on the impact of discharges in this paper.

Inpatient bed LOS can vary significantly from patient to patient. In addition to regular inpatients (27,000 in our one year data), there are two separate categories of patients called "day-stay" and "observation patients". Day-stay patients, as the name suggests, are patients who undergo small procedures like tonsillectomy and stay for 24 hours or less in an inpatient bed (ASCA [2013]). Observation patients refer to those patients whose conditions can be treated in 48 hours or less, or when the cause for the symptoms has not yet been determined. Some examples are nausea, vomiting, and some types of chest pain. Bed requests for these patients are typically made through the ED (CMS [2011]). In the data we analyzed, day-stays and observation patients sum up to 20,000 patients. Thus, in total with regular inpatients, we have a total of 47,000 total patients who used an inpatient bed for the one year period of interest.

Regular inpatient bed requests can get admitted through the ED, surgical units (this includes elective surgeries such as hip and knee replacements as well as emer-

gency surgeries), from physician offices (direct admits), from other community hospitals (transfers from other hospitals). Categorizing patients by these admit sources and their MDC, mimics accurately which units they get admitted to and how long they stay in an inpatient bed. Figure 4.4 shows total annual bed requests of the major sources with respect to hours of the day. We can clearly see the time-varying arrival nature of each source.

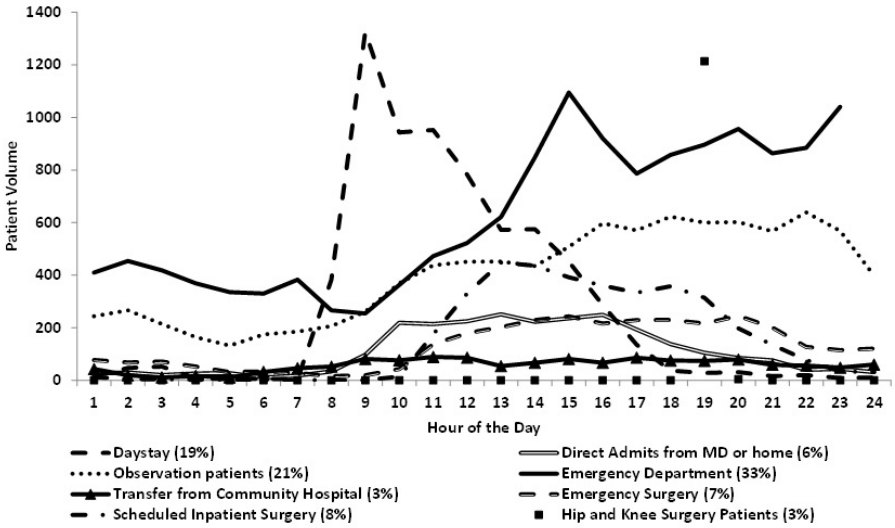


Figure 4.4: Arrival pattern by patient sources

Figure 4.5 shows the LOS and daily bed request rate for the patient sources discussed above. Day-stay patients who exhibit a high annual volume of 10000 patients, spend less than a day in the hospital. On the other hand, ED patients present an annual volume of 14000 as well as a high LOS of around 5 days. The "controllable" patients – elective surgery patients, who can be scheduled in advance – are denoted with solid fill, whereas the "uncontrollable" sources (patients from the ED, for example) are represented with solid diamonds, and the horizontal lines are

for patients somewhere in between. Calculating the "patient bed days" (total volume * average LOS) consumed by each patient category, suggests that ED patients, at this hospital at least, consume the majority of inpatient capacity. Even though the literature about changing surgical schedules is abundant (Helm and Van Oyen [2010], Bekker and Koeleman [2010]), for this specific hospital the impact of ED patients dominates all the controllable sources. Note however that surgery rates because they are elective are scheduled over 5 weekdays, whereas emergency surgeries are admitted throughout the whole week (both weekdays and weekends). Additional analysis on elective surgeries can be seen in Appendix B.

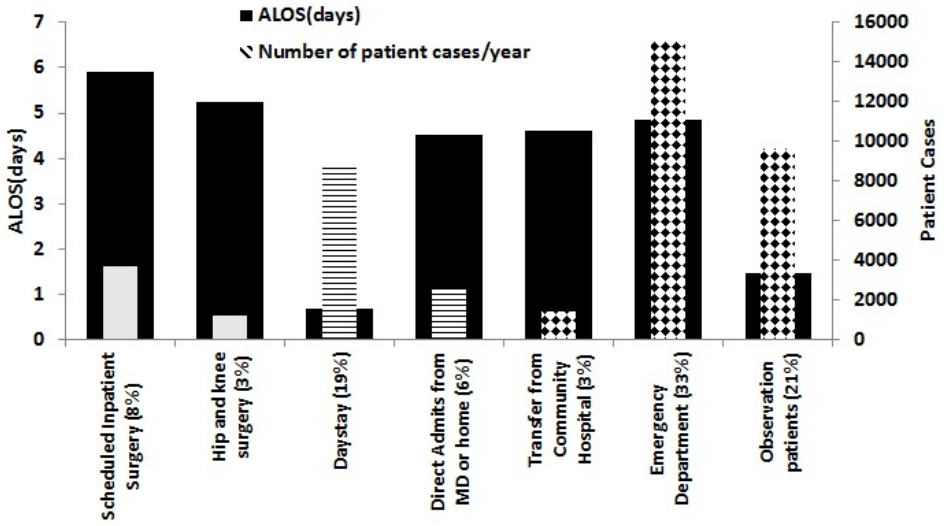


Figure 4.5: Volume and LOS values of patient sources

There are 25 departments that the patients get admitted to, which total to 575 inpatient beds. The medical specialties include: adult respiratory, oncology, day-stay, medical, observation, congestive heart failure (CHF), surgical, interventional, critical care, women health, renal, neurology, orthopedic and pediatric medical and

surgery units. We have analyzed the units in terms of their bed capacity, daily arrival rate, the mean and variance of LOS, utilization level and average percentage of total discharges before noon.

Table 4.3 provides twelve of the most highly utilized units. Here we define "utilization" with an aggregate simple formula: daily bed request times the average LOS, divided by the number of beds. The LOS values exhibit significant variability, in most cases it is higher than the mean. There are also some hospital specific dynamics that affect the hospital-wide flow. Different units host different kinds of patients. The highest utilized unit S2, a medicine-telemetry unit, generally hosts "socially challenging" patients (like overdose patients who require further care). Telemetry service is often recommended after a heart attack, or when a patient is seriously ill. Nursing staff in telemetry units is usually highly trained so that they can respond to emergent issues quickly (McMahon [2014]). Another unit to point out is APTU, the psychiatric unit which has the highest LOS and highest variability. In both of these units the predictability of discharge is harder to estimate, than a surgical ward, because these patients generally require a post-acute care service. Hospital specific subtleties like these affect the whole admission and discharge process and by using random sampling from the historical data we are able to incorporate these factors implicitly to our model.

Table 4.3: Unit specific analysis

Unit	LOS (days)	Std Dev LOS	Daily Rate	Capacity	Utilization	% discharge before noon
Medical Telemetry	5.09	5.85	5.16	26	101%	16%
Cardiac CHF	4.19	4.29	7.58	32	99%	24%
Cardiac interventional	2.64	3.27	11.87	32	98%	23%
Neurological	3.91	5.08	9.94	41	95%	13%
Renal	3.25	4.19	6.92	24	94%	27%
Adolescents	2.6	3.38	2.86	8	93%	27%
Medical Respiratory	5.27	7.35	5.36	31	91%	29%
Surgical/Orthopedic	4.74	5.02	6.47	34	90%	15%
General Medical	3.26	3.21	11.95	44	89%	23%
Psychiatric	8.68	10.87	2.79	28	87%	21%
Intermediate Surgical	5.41	6.25	6.97	44	86%	9%
Short Stay Surgical	4.25	5.61	6.28	32	83%	12%

4.5 Simulation Model and Analysis

Figure 4.2 and pseudocode presented in Appendix C, show the main idea behind our simulation model. There are M inpatient bed request sources. The number of requests from source i in hour t is denoted by the random variable $\lambda_{i,t}$ and is sampled randomly without replacement in order to reflect the time of day and day of week effect. These requests fall into some MDC category and are consequently mapped into demand for N inpatient units. The total number of bed requests for unit j at hour t is denoted by the random variable $\lambda'_{j,t}$. Each unit has B_j beds, and each unit is a time-varying $G/G/B_j$ queue. The arrival rate in each hour follows some general stochastic process; Poisson arrival rates are not a bad assumption, but in our case, we use arrivals sampled from historical data, hence "G" in the queuing notation. The random variable LOS_j indicates the service time in unit j and follows

some general distribution. We provide examples for arrival rates and LOS values for different admission sources that were used in sampling in Appendix B.

Some patients make their way to an inpatient unit via the ICU where critical care is provided. We assume that these patients spend a deterministic amount of time specific to an MDC, $CritLOS_{MDC}$, before requesting for a regular inpatient bed.

Each hour, bed requests are fulfilled on a first come first served (FCFS) basis. The patients are ready to be discharged from the hospital after their LOS is completed. They join a discharge queue, which has a capacity of D_t in hour t. To start with, patients are discharged on a FCFS basis as well; so there is no speeding up or slowing down, which is commonly observed in practice (Jaeker et al. [2012], Kc and Terwiesch [2009]). As an alternative discharge policy, we also consider prioritizing discharges in units which have the longest admission queues. The bed is available after the bed turnover time (a deterministic value) is complete.

In each unit j, a queue Q_j develops consisting of those patients waiting for a bed to become available. We assume that the patients simply wait until they receive a bed, irrespective of the size of the queue; i.e. there is no balking. In practice, the hospital may use alternate strategies, such as using free beds in other units, though this is not desirable. Note that the queue is not a physical waiting line of patients; rather it consists of patients waiting in different parts of the hospital (ED, PACU) or other hospitals. Waiting time measured as the time difference between in the bed time for the right bed and time of bed request. So this measure includes ED boarding, PACU holds, and all other waiting times relevant for an inpatient to be placed in an inpatient bed. It is also possible that a patient is waiting at home for an

inpatient bed, after a bed request was made by a community physician the patient consulted with.

We used C# for the simulation representing the whole hospital-wide flow. We run the model for a year, with hourly increments, kept the warm-up period as 2 months. We sample from historical data, observed unit requests and LOS values. All the $\lambda_{i,t}$, $\lambda'_{j,t}$ and LOS_j for each hour and day of week are sampled randomly. In the sampling process, we retain time of day and day of week effects for arrivals. As an example, for Monday 8 AM, we randomly sample, without replacement, from arrival, MDC and desired unit requests observed on 52 Mondays at that exact hour. We also develop a simulation model in Arena for internal validation purposes (we provide a detailed explanation of the Arena model in Appendix D).

4.5.1 Replications

We have compared the waiting times and number of people in the queue, using various discharge profiles. In order to have an unbiased comparison, we use the common set of random patients for each replication. This is the common random numbers (CRN) approach which serves as a variance reduction technique when comparing different policies (Banks et al. [2004]). We have used 10 replications, following Shi, P. and Chou, M. C. and Dai,J.G. and Ding, D. and Sim, J. [2012] who also use 10 replications to perform their analysis. More replications will lead to a higher accuracy, however, due to the computational complexity (around 2 hours for each run) we limit the number of replications. Also, the main motivation of these runs is to be able to

compare and analyze the impact of using different discharge policies for improving the bed capacity planning.

4.6 Analyzing the Impact of Discharge Policies

Recall that the purpose of this paper is to test different discharge profiles by changing the D_t and observing the impact on queue lengths Q_j and the waiting time. We are evaluating 3 components of discharge profiles: (1) Discharge windows which determines the hours of the day when the discharges are allowed; (2) The maximum capacity for discharges in each hour of the discharge window (the D_t values); and (3) The prioritization of discharges in each hour based on admission unit queues (i.e. which patients should have first access to discharge capacity in a given hour). We evaluate different combinations of these 3 components and compare it with the baseline which represents our partner hospital's discharge operations. We now present all the discharge profiles we test in our simulation and also provide the rationale for each.

4.6.1 Baseline

The baseline discharge profile for BMC was briefly described in the Introduction. The discharge window is currently from 10 AM-7 PM. The hourly maximum discharge capacity is set to the average number of discharges achieved by the hospital in the one year period studied. As we explained earlier, discharges follow a bell curve that peaks between 2 and 4 PM. Starting with the hour 10-11 AM, we set D_t equal to 5, 7, 11, 12, 14, 18, 16, 10, 6, 5 until 7 PM. For all other hours D_t is 0. Currently

at the hospital, there is no obvious prioritization of discharges, so we assume in the baseline that discharges are done on a FCFS basis. In other words, a patient whose LOS finishes earlier will be given priority, irrespective of which unit the patient is from.

4.6.2 DP2: Maximum capacity of 10 in each hour from 10 AM-7 PM, no prioritization

We analyze restricting the number of discharges to 10 each hour. Notice that in the baseline, the hospital achieves up to 17 discharges on average in each hour. Therefore 10 is a very reasonable upper limit and was suggested by our collaborators. Thus, in this policy D_t is restricted to be 10 in each hour from 10 AM-7 PM; for all other hours D_t is 0. This promotes a more even or uniform discharge workload for the hospital staff in the window rather than having a peak in the afternoon. Discharges are carried out on a FCFS basis (no prioritization).

4.6.3 DP3: Early in the day discharge policy, 10 AM-7 PM, no prioritization

The main motivation of early in the day discharge (EITD) policy is to align the discharges and the admissions by pushing some of the afternoon discharges to the mornings so that the beds are available before the demand builds up. In this discharge profile, D_t is only restricted to be less than the remaining number of average daily discharges. Because of this, most of the patients leave the hospital in the first 2 hours of the window (10 AM-noon). These patients have already completed their LOS overnight and have been waiting for the hospitalist to discharge them; hence

the name early in the day discharge. Discharges are carried out on a FCFS basis (no prioritization). The actual number of discharges realized each hour is analyzed in Section 4.7.2.3.

This profile is of particular interest since many hospitals have been emphasizing that discharges should happen before noon. This is quite a difficult process because even if the patients are ready, their post-hospitalization transition (family pick-ups, rehab facility and so forth) may not have been coordinated. However some of the patients are more amenable to early discharge, especially "simple discharges" that account to 80% of hospitals' discharges. These are the patients who are discharged to their homes or do not require complex planning, like most of the surgical floor patients (DH [2004]). Motivated from this, Department of Health in UK has reported a 40% decrease in the number of elective surgery cancellations in Nottingham City Hospital, simply by implementing a policy based on discharging medically fit patients by midday.

4.6.4 Expanded discharge windows

Our collaborators in BMC also urged us to test the feasibility of expanding discharge hours as an alternative to early in day discharges, because they felt that discharges by noon were very difficult to implement in practice (as explained in Section 4.6.3). So instead of having a 10 AM to 7 PM discharge window, an expanded window from 10 AM to 9 PM or 10 AM to 11 PM could be tested. Each hour in the expanded window $D_t = 10$ and 0 otherwise. We test three discharge profiles with expanded windows:

DP3: Maximum capacity of 10 in each hour from 10 AM-9 PM, no prioritization.

DP4: Maximum capacity of 10 in each hour from 10 AM-11 PM, no prioritization.

DP5: Early in the day combined with a 4-hour expanded window, 10 AM-11 PM, no prioritization.

The end result of an expanded discharge window is that more patients could be discharged in the day; more beds become available the next day as a result. The expanded window is also more in line with the hospitalists' natural prioritization rules. They can see the most recently admitted patients, who need more urgent attention, in the morning, and get to the patients who are ready to be discharged later in the day. Expanding the discharge hours also allows them to discharge those patients who would unnecessarily wait until the next day. The families of patients may be more available to pick up patients in the evening rather than during the day. The actual number of discharges realized in each hour after 7 PM is analyzed in Section 4.7.2.3.

Caveats do apply. An expanded discharge window does require staggering of shifts so that hospitalists are available between 7-9 PM or 7-11 PM (like nurse shifts). Additionally ancillary services that are essential for a patient's discharge process also need to be available in the evening hours. The patients need to pick up their medication from the pharmacy, and perhaps equipment such as walkers. Patient transport and valet services are also needed to escort the patients out of the hospital. In general, the more services that patients need after their discharge, the greater the staff availability needs to be in the evening hours. Thus the hospital needs to adopt a case-by-case approach, and utilize evening discharges wisely. Our partner

collaborators in BMC agreed that these changes that need to accompany expanded hours are indeed feasible.

4.6.5 Prioritization of discharges

Up until now, decisions in our simulation have not been responsive to the state of the system. Thus even when the hospital is facing a gridlock, our assumption is that the hospital carries out its regular operations and queues continue to grow. However, in practice hospitals may respond by canceling elective surgeries, diverting ambulances, or by speeding up discharges. All of these have potentially negative outcomes.

We take a different approach to model the hospital's responsiveness. In our simulation, we prioritize the use of hourly discharge capacity. This prioritization is based on front-end admission queues for each unit. If queue is larger than some threshold the hospitalist and related staff first focus on discharging patients from these units. However, it is important to point out that these are not hasty discharges (which may cause readmissions), rather a policy that allocates the restricted discharge capacity to the units that require it the most. Using the red-yellow-green system discussed in (Resar et al. [2011]), we categorize the units into two: red and green units. For the red units the current queue length of the unit exceeds a predefined threshold. Green units are those that do not exceed this value. Prioritization in our model implies that the hourly discharge capacity should be first used for the red units. Thus, this state-dependent discharge policy observes the congestion in the first service line (admitting patients to an inpatient unit) and accordingly adjusts the use of capacity in the second service line (discharging patients from the unit). Pseudo-code for priori-

tization is provided in the Appendix C. We test this prioritization for the following discharge profiles:

DP6: Baseline with prioritization.

DP7: Maximum capacity of 10 in each hour from 10 AM-7 PM with prioritization.

DP8: Maximum capacity of 10 in each hour from 10 AM-9 PM with prioritization.

4.6.6 DP9: 24-hour discharge

This discharge profile cannot be realized in practice and is meant purely as a benchmark. D_t is unrestricted in each hour of the day. No prioritization is necessary as patients can leave as soon as their LOS is finished.

4.7 Results of the Simulation

4.7.1 Validation

Before trying to improve the existing system, validation was the initial step. The validation involved two steps: stakeholder face validation and comparison of means of the inputs and outputs (as discussed in Montgomery and Davis [2013]).

Using mathematical models to solve problems in clinical settings is a very complex process. The assumptions supporting the mathematical model need to be clinically realistic. Tucker et al. [2001] remind us that clinicians make decisions based on their perceived patient priorities, rather than system efficiency. These decisions dictate clinicians' actions (prioritizing which patients to see first). The clinician is motivated

by their perceived action value to patient care, and system efficiency is secondary to this goal. The assumption underlying mathematical projections only include value weights programmed into the model. This project involved the interactive face to face process of reviewing and comparing mathematical assumptions and the clinical assumptions. This process is time consuming but essential to validate the model.

Thus, face validation is a result of our close collaboration with our team of clinicians and data managers. It was an iterative process and we asked questions like "Does it represent the reality?", "What should be changed?" and so on. We have also discussed with our clinical collaborators about the system dynamics like queue sizes (unfortunately we cannot validate this precisely with the data-set). On average at any given hour 40 people waiting to be admitted to an inpatient unit, was an accurate estimate to the queues in our partner hospital. We have performed sensitivity analysis (like changing capacities in different units) in order to further validate the results of our simulation model.

After face validation, we have also compared the means and quantiles in our model, with the empirical distribution. We have compared input and output variables with the empirical data, including the comparison of: waiting times, admission patterns and LOS values for patients on MDC levels and utilization levels for units.

Some of the output variables overestimate the empirical values. There are 3 main reasons for the overestimated values of our simulation model: firstly in our simulation model we do not model redirections between units, while in the hospital patients would be overflown to other units. The simulation model mimics a perfect world, in which patients are only admitted to their primary unit. Secondly, we sample

from biased LOS values; since these values from the data-set already have embedded delays and non-value added times. Lastly, our simulation model is not responsive to over-crowdedness, whereas in real life the hospital would go on "code red" to cope with the congestion by employing policies like ambulance diversions, cancelling transfers, elective surgeries and so on. Thus, in some sense we are modeling a worst case scenario.

Even though the values do not match precisely, our main objective is to compare different scenarios and policies to improve the patient flow. Also, even if the waiting times or queue sizes are not precisely the same, the congestion pattern is the same. So the most congested units are the same, the same is true for the patients who wait the most.

We observe the phenomenon as pointed out in Green [2012], that a hospital may have ample beds in some units and insufficient in others, resulting in long ED waits and ambulance diversions. This is simply because, not all the beds are identical. Thus, needed bed capacity is highly dependent on the patient mix. Green also mentions that the smaller the system, the longer the delays will be for a given utilization level; and the greater the variability in service times, the longer the delays at any utilization level. So the smaller units with higher variability in LOS will have a higher wait time. This can also be seen from our results as well. For instance, psychiatric unit with the highest variability in LOS experiences long queues.

4.7.2 Impact of discharge policies

We now present results of 10 simulation replications for the various discharge profiles, with and without prioritization. We begin with the analysis for the average queue size. The average queue size represents the average number of people waiting to be admitted to a unit waiting in ED or PACU, or in the community hospitals. We use one-factor ANOVA to analyze the differences in average queue size between the discharge profiles (Figure 4.6), and also the all-pairs Tukey test (as can be seen in Table 4.4).

The red lines in Figure 4.6 represent the quantiles with the box plot, the blue lines the standard deviation, the green horizontal bar represents the mean for each category, and the top and bottom of the diamond shape are the 95% confidence intervals. Lastly, the horizontal line is the overall mean queue length across all discharge policies and replications. The discharge profiles are presented in descending order in terms of the average queue size observed. DP9 represents the 24 hour discharge policy, which is a hypothetical best-case benchmark; DP6, DP7, DP8 are the prioritized discharge policies, the rest are the un-prioritized discharge profiles and the baseline represents 10 AM-7 PM with empirically observed discharge capacities (see Section 4.7.2.3).

The connecting letters report in Table 4.4 summarizes the results of the all-pairs Tukey tests. If two discharge profiles share the same letter, they cannot be said to be statistically significant. However, statistical significance, while important to acknowledge, should not be confused with clinical significance. Clinical significance has a qualitative component; in our case, it is decided by our clinical collaborators.

For example, a 6-person average reduction in queue size is non-trivial even though it may not be statistically significant. In our results we find that a discharge profile when compared to another (1) may not be either statistically or clinically significant; (2) may be clinically significant, but not statistically significant; and (3) may be both statistically and clinically significant. Table 4.4 suggests that the third type of conclusion is prevalent only with regard to the prioritized discharges.

Table 4.4: Connecting letters report for queue size

Level					**Mean**
Baseline	A				45.849
DP1	A				46.122
DP2	A	B			39.570
DP3	A	B	C		38.841
DP4	A	B	C	D	33.733
DP5		B	C	D	31.319
DP6		B	C	D	28.596
DP7		B	C	D	27.012
DP8			C	D	24.673
DP9				D	20.481

Our results can be summarized as follows:

(1) The empirical discharge distribution (Baseline) is neither statistically nor clinically different from a discharge profile that restricts the number of discharges to 10 each hour (DP1). Thus allowing a steady discharge rate of 10 every hour is not different from a discharge policy that peaks in the afternoon.

(2) If the majority of discharges happen before noon, as in the early in the day discharge policy (DP2), then there are 6 less people waiting in the queue compared to Baseline (clinical significance) but there is no statistical difference. This supports the findings in Shi, P. and Chou, M. C. and Dai,J.G. and Ding, D. and Sim, J.

[2012]. The exact number of people discharged before noon is provided in Section 4.7.2.3; which will demonstrate that early in the day discharges will be very difficult to implement in practice. In DP2 the discharge capacities each hour between 10 AM-7 PM are only restricted to be less than the remaining number of daily discharges. Despite this, queue sizes do not decrease significantly. This suggests that not enough patients' LOS ends in the 10 AM-7 PM window.

(3) With 2 additional hours of discharge and a steady maximum capacity of 10 discharges in each hour (10 AM-9 PM, DP3), we have 7 fewer patients waiting compared to Baseline (clinically significant). Notice that DP3 matches the performance of early in the day discharge (DP2). Thus expanding discharge by two hours while limiting the maximum hourly discharge capacity to 10 produces the same effect as performing a large (and impractical) number of early in the day discharges. In fact, section 4.7.2.3, we will show that the number of patients discharged in the hours between 7 PM and 9 PM is actually well below 10 for each hour.

(4) With 4 additional hours of discharge, and a steady maximum capacity of 10 discharges in each hour (10 AM-11 PM, DP4) there are 12 fewer patients waiting compared to Baseline (clinically quite significant, but not statistically). This reinforces the idea that expanding discharge windows while keeping a practically feasible and steady limit on discharge capacity has a stronger impact than allowing an early in the day discharge policy between 10 AM-7 PM.

(5) When early in the day discharge profile is combined with a 4-hour expanded discharge window, we have DP5. Such a discharge profile is not realistic since it requires too much alteration of current practices; nevertheless, DP5 serves as a

benchmark. We see that DP5, while statistically different from Baseline and DP1, produces only a 2.5 patient reduction in queue size compared to DP4 (not statistically significant and perhaps not clinically significant either). This again suggests that expanding the discharge windows has a stronger effect than carrying out early discharges between 10 AM-7 PM.

(6) We begin to see both statistical and clinical differences when discharges are prioritized in units that have the most patients waiting (DP6, DP7 and DP8). We see also from Figure 4.6 that the higher percentiles of the average queue size (for each discharge profile there are 10 average queue size observations obtained from the 10 replications) are also reduced drastically. Using the current or empirically observed discharge capacity as the maximum capacity for each hour and a discharge window of 10 AM-7 PM with prioritization, produces a statistical improvement from the Baseline: it leads to 17 less patients waiting to be admitted. The only difference between Baseline and DP6 is prioritization: the only change in practice is that each hour the hospital staff (physicians, case-managers, nurses, valets and escorts) has to prioritize their discharge activities in units that have longer front-end (admission) queues. Notice also, that DP6 produces a greater improvement (though not statistically significant) than using the combination of early in the day discharge policy with a 4-hour expanded discharge window (DP5). DP7 shows identical results as DP6.

(7) The impact of prioritization is further enhanced under when the discharge window is expanded by 2 hours (DP8), and a maximum of 10 discharges are allowed each hour. Now, we have 20 fewer patients waiting (47% improvement) to be

admitted, compared to Baseline. This difference is both clinically and statistically significant. Indeed, DP8 is comparable to the queue size observed from 24 hour discharge policy (DP9). DP9 is only a benchmark – a lower bound that can never be achieved. It is surprising how close DP8, which has some feasibility in practice, performs with regard to this benchmark.

4.7.2.1 Unit specific analysis

We present how the queue size changes with different discharge profiles, in Table 4.5, for the 5 units with the highest queues; Medical Telemetry, Renal, Medical Respiratory, Cardiac Interventional units. As can be seen prioritization mostly benefits Medical Telemetry unit, whereas the queue size in the Neurological unit is worse off with this policy.

Table 4.5: Average queue size

Admit unit	**Baseline**	DP1	DP2	DP3	DP4	DP5	DP6	DP7	DP8	DP9
Medical Telemetry	16.46	16.31	14.81	13.37	16.36	13.08	9.75	9.78	8.98	8.01
Renal	4.63	4.44	3.86	3.41	4.43	3.42	2.95	3.00	2.53	2.26
Medical Respiratory	3.81	3.69	3.37	3.08	3.67	3.02	1.77	1.81	1.67	2.15
Cardiac interventional	4.40	4.62	3.00	2.18	2.27	1.56	1.66	2.13	1.53	1.00
Neurological	3.36	4.06	2.62	1.85	1.82	1.30	2.02	2.44	1.76	0.78
SUM	32.66	33.12	27.67	23.90	28.56	22.38	18.15	19.15	16.46	14.21
% improvement		**-1%**	**11%**	**19%**	**9%**	**22%**	**32%**	**30%**	**35%**	**40%**

We also study the queue size quartiles for the two units of interest: Medical Telemetry and Neurological unit in Table 4.6. Note that average queue sizes are highly driven by higher percentiles, with median and 25th percentile typically having a value of 0 for most of the units. Only Medical Telemetry unit has a queue size greater than 0 in the first quartile, observed in the Baseline discharge profile.

Table 4.6: Queue size percentiles using different discharge policies

	BASELINE		EITD		PRIORITY	
	Med-tele	**Neuro**	**Med-tele**	**Neuro**	**Med-tele**	**Neuro**
1st quartile	3	0	3	0	1	0
Median	14	0	13	0	8	0
3rd quartile	27	2	26	0	15	0
Avg	16.85	2.66	16.76	1.86	9.92	1.55
Max	55	34	55	29	39	24

4.7.2.2 Waiting time analysis

The second output measure of interest is the waiting time, which is a weighted average of the admissions waiting time. Different from previous research in the literature, it is not only based on ED boarding time, but rather includes the PACU holds, transfer waiting times and ICU holds as well. The reason why the waiting times are higher than the average values in the literature is because we are calculating the time for patients to be admitted into their primary units and consider the waiting times from all different patient sources. The improvements in waiting times follow the same trend as queue sizes, as can be observed from the ANOVA analysis in Figure 4.7. The confidence intervals for the waiting times are provided as well.

4.7.2.3 Resulting discharge capacities

In order to analyze the resulting discharge capacities from different discharge profiles, we have looked at two of the most highly utilized units: Medical Telemetry and Surgical/Orthopedic. The limited discharge capacity is especially allocated for these units under the prioritized discharge policy (Figure 4.8). Medical Telemetry unit benefits the most from this prioritization, in improving the long waiting times.

We have also analyzed these different discharge profiles, in terms of the actual realization of hourly discharges for the whole hospital. The hourly discharge capacity thresholds and the actual discharge number, using different discharge profiles can be observed from the Figure 4.9 below.

The infeasibility of early in the day discharge policy can clearly be explained with the graph. As can be seen the first two hours together require almost 70 discharges, which is more than the double of the average discharge capacity observed in peak hours (approximately 16 patients). And even with this unlimited discharge capacity, the improvements are not significant.

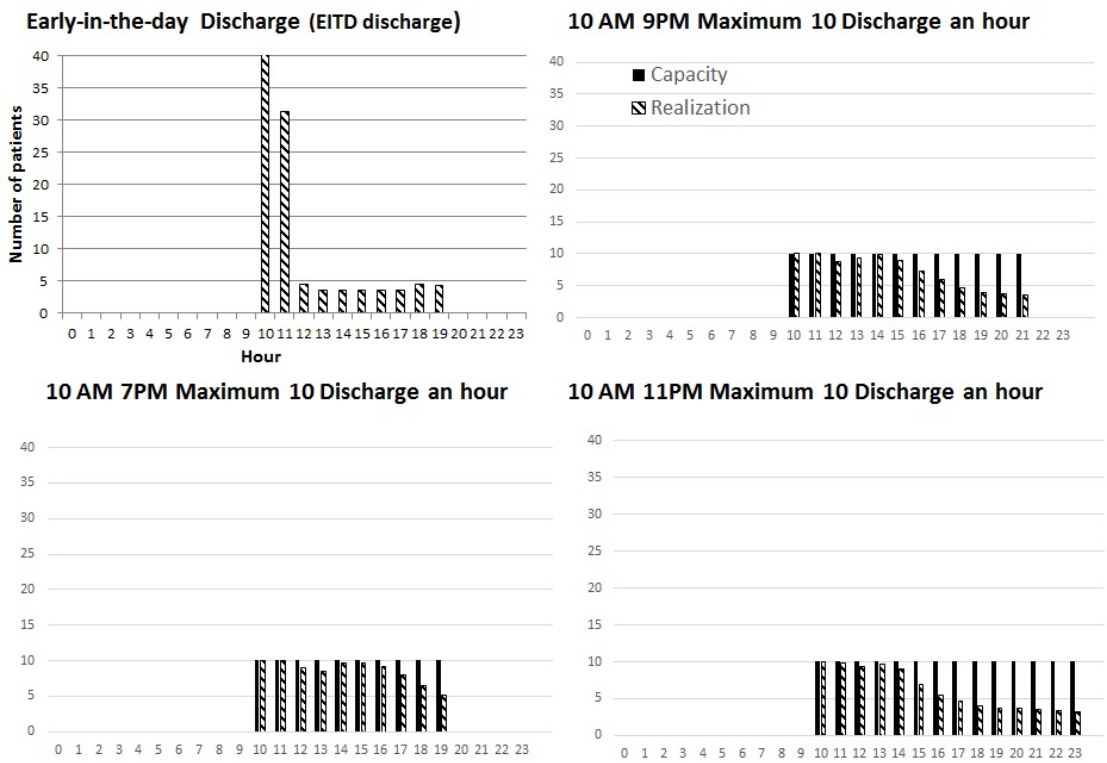


Figure 4.9: Capacity thresholds and actual realizations

It is also noteworthy that with only 7 to 13 more discharges after 7 PM the queue sizes and waiting times improve drastically with the expanded discharge window. The reason is that the discharges are performed more evenly throughout the day. The discharge capacities are not reached in most of the cases, as can be observed from Figure 4.9.

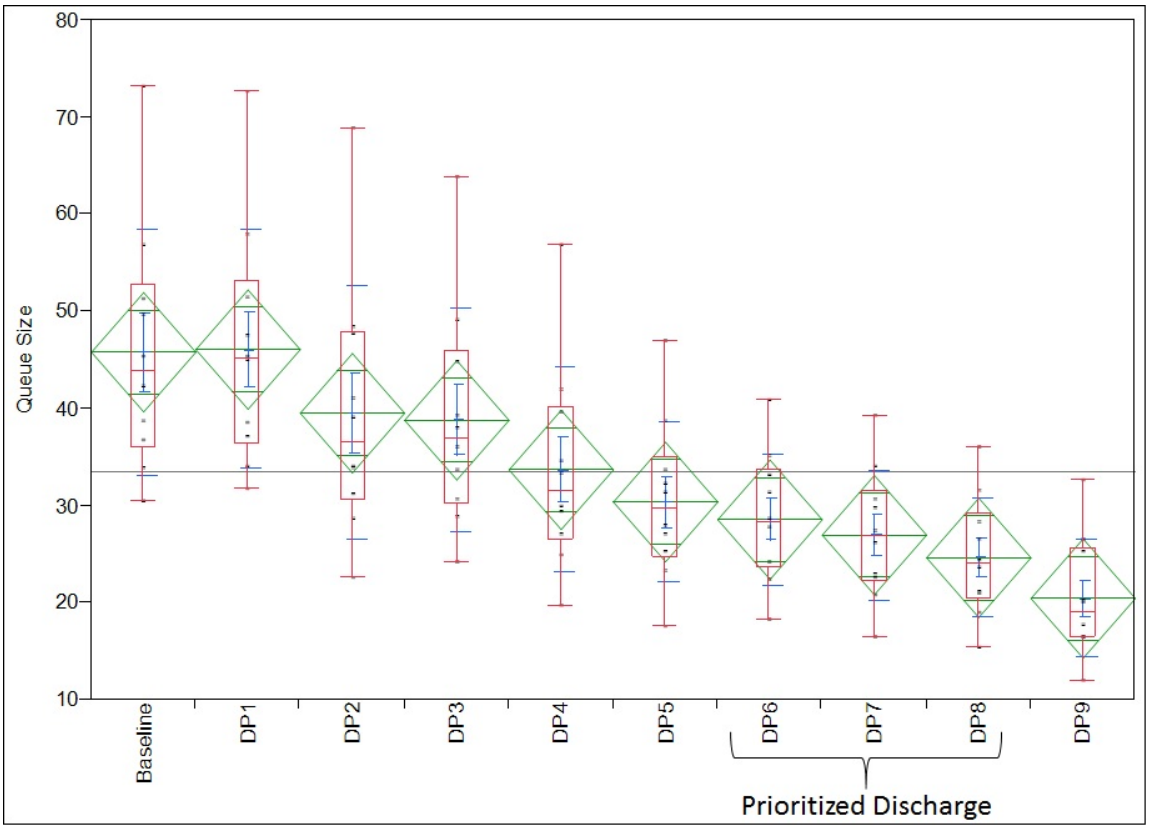


Figure 4.6: Queue size ANOVA analysis where Baseline: 10 AM to 7 PM empirical discharge distribution, DP1: 10 AM-7 PM max 10, DP2: 10 AM-7 PM EITD, DP3: 10 AM-9 PM max 10, DP4: 10 AM-11 PM max 10, DP5: 10 AM-11 PM EITD, DP6: 10 AM-7 PM Empirical Priority, DP7: 10 AM-7 PM max 10 Priority, DP8: 10 AM-9 PM max 10 Priority, DP9: 24 hour discharge

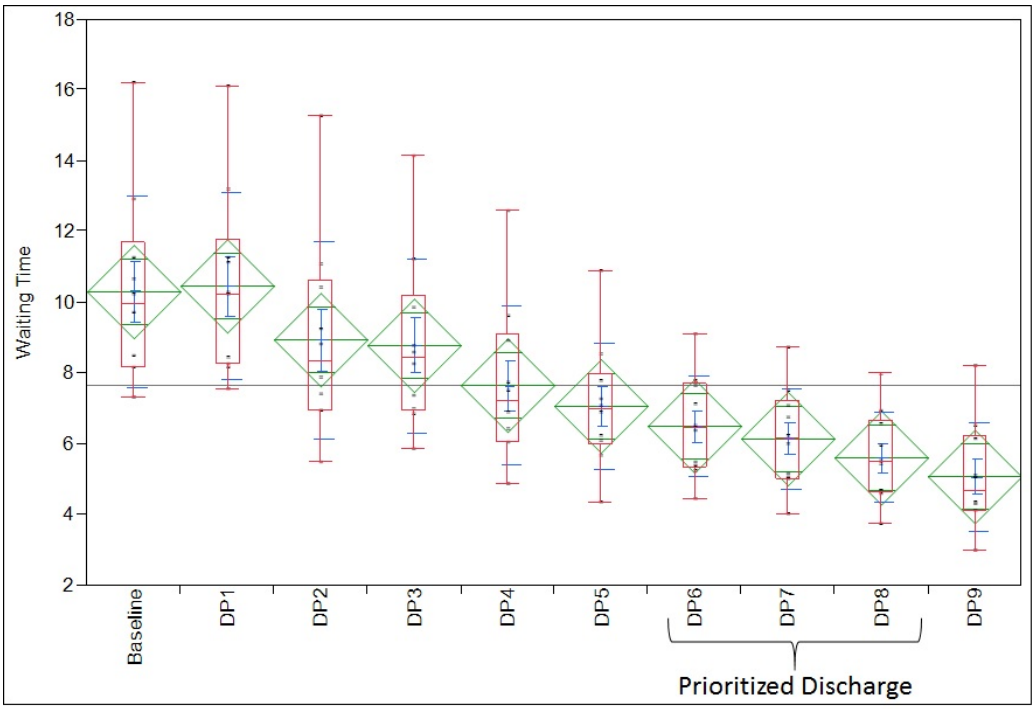


Figure 4.7: Waiting time ANOVA analysis, where Baseline: 10 AM to 7 PM empirical discharge distribution, DP1: 10 AM-7 PM max 10, DP2: 10 AM-7 PM EITD, DP3: 10 AM-9 PM max 10, DP4: 10 AM-11 PM max 10, DP5: 10 AM-11 PM EITD, DP6: 10 AM-7 PM Empirical Priority, DP7: 10 AM-7 PM max 10 Priority, DP8: 10 AM-9 PM max 10 Priority, DP9: 24 hour discharge

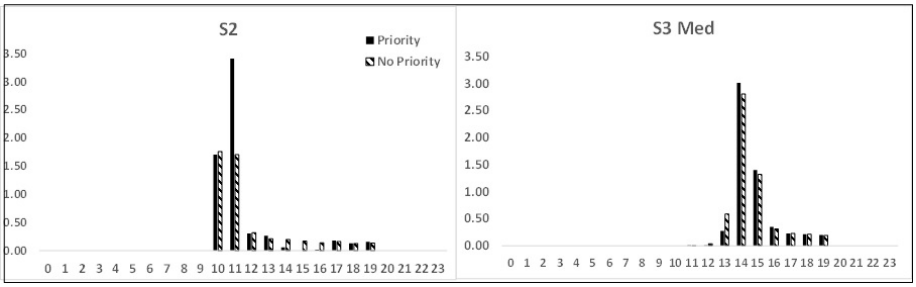


Figure 4.8: Discharge profile for two units Medical Telemetry and Surgical/ Orthopedic

4.8 Discussion and Conclusion

As discussed in Powell et al. [2012] effective solutions require a system-wide approach. Thus, we provide a hospital-wide patient flow model as opposed to a unit specific analysis, which can be used a decision support mechanism. The clinical leaders involved in this simulation used the results in deciding on the capacity for new units. For instance, the medical telemetry unit that was identified as having the longest queue, was moved to a larger unit with more beds.

In our project, gridlock and long wait times for inpatient beds were compelling issues to administrators, and managers. Clinicians (physicians and nurses), however, are more influenced by their most critical patients' needs. Realistic time-frames and goals used in the model need to reflect this tension between priorities. The early in the day discharge option is an example of the conflict between the individual clinicians' decisions and the management targets for this model (See Table 4.1). Historically (see Figure 4.1), the maximum number of discharges accomplished in an hour has been 17. If the early in the day discharge model is used – there would need to be 32 to 40 discharges per hour (Figure 4.9), demonstrating an unrealistic clinical target, with the same system rules and resources. Hospital administrators have been recommending these early morning discharges. However, in 2004, Kealey and Asplin [2004] reported the best practice "Forget about trying to get all discharges out by 11 AM", instead they recommend scheduled discharges.

Many methods have been investigated in the literature to alleviate the bed congestions including using flexible beds, increasing the number of beds (Green [2003]), optimizing the surgical schedule (Bekker and Koeleman [2010]) or creating some

kind of admissions control mechanism (Helm et al. [2011]). In our research we investigate using effective and realistic discharge policies, that could be implemented and actually create a significant improvement. We evaluate various discharge policies including expanding discharge windows, limiting the number of discharges to a threshold and prioritizing discharges based on the admissions queue. We conclude that effective discharge policies have a significant impact on reducing waiting times. For example, expanding the discharge windows only by a few hours provides substantial benefit, although not as significant as prioritized discharge policies (which reduces queue sizes up to 48%).

By exploring other "windows" (evening hours) for discharge that could decrease the wait time and queue size, the mathematical model gives the clinicians "new eyes" to explore a new model to use increased discharges to decrease congestion. Clinical administrators underestimate the conflict of priorities between system efficiency and clinical priorities. Engineers work to identify the mathematical system possibilities and project the impact of system changes. The best solution exists at the intersection of all three partners in this modeling, as in Figure 4.10.

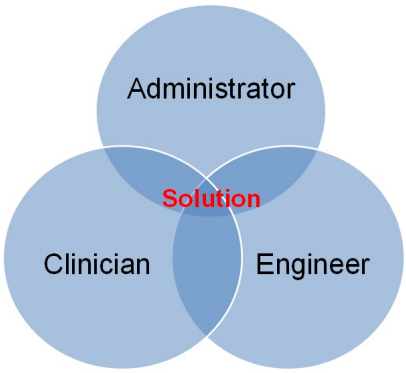


Figure 4.10: Best solution

Like any model there are some points that we have failed to address in this paper. Our model can be potentially extended in several directions. First of all, we assume a dedicated system in our simulation for bed capacity planning in which all of the patients get admitted to their primary units. This is in fact desirable both from the hospitals' and patients' point of view. However, in reality most of the time, the patients will be redirected to other units (non-primary units) after a certain time. This will decrease the waiting times and queue sizes, however the patients will not be receiving the exact medical attention that they require. Even if the room is equipped adequately, nurses are best prepared to care for the diagnoses commonly accepted to their units. Each unit has specialty protocols, common treatments, known by the unit nurses. When they care for patients with different diagnoses and issues, the match between the patient and nurse will not be optimal. This mismatch may result in delayed or inappropriate care. Furthermore, the patients in their non-primary units will then have to be re-transferred to their primary requested unit. This results in unnecessary costs, bed turnovers, potential health risks related with unsuitable admit unit and patient dissatisfaction resulting from an unnecessary transfer (West [2010b], West [2010a]). Thus in this study, we choose not to incorporate transfers, in order to model the worst case scenario. As an extension to this study, we have integrated overflow transfers to our simulation model and performed some preliminary analysis on the impact of these (see Section 4.10 for detailed explanation).

An important point to consider, like in any data driven modeling, is the reliability of data. Electronic health record generated admissions data consist of numerous

time-stamps, the accuracy of which relies on human input. However, we have done numerous data analysis and checks with our collaborators to validate the accuracy.

4.9 Ongoing Research

The current research on patient flow modeling can be potentially extended in the following directions:

- The number of discharges from each admit unit can be limited to a certain threshold every hour. This would result in a more realistic model, since the data analysis has pointed out that there are at most two discharges that can happen in any unit any hour. Incorporating this to the simulation model is essential in order to simulate the hospital-wide flow more accurately. Our preliminary analysis has shown that this does not significantly or statistically impact the queue sizes, when discharges are prioritized with these unit level constraints in mind. The initial results for the one replication is provided in Appendix E.

- Instead of prioritizing some units, a prioritization scheme for some of the patients can be explored as well. Certain types of patients might have more impact on alleviating bed congestions than others. For instance, our nursing collaborators have hypothesized that if observation patients, who face a more problematic discharge process, are prioritized over inpatients this might lead to significant improvements in ED congestions.

- We have assumed that the bed turnover times are deterministic. However, in the afternoons when the majority of discharges happen, cleaning time typically takes longer than the average (around 47 minutes for our collaborating hospital). In the simulation model, these durations can be modeled as a random input parameter depending on the housekeeping workload, or at least a value that varies over the hours of the day.

4.10 Incorporating Transfer Activities Among Different Units

On average 40–70% of inpatients in U.S. hospitals are transferred each day, thus patient transfers are an important part of hospital patient flow (Abraham and Reddy [2010]). There are two main types of patient transfers in a hospital; either it may be medically necessary for the complex patients to receive medical treatment from different units during their LOS (Hilligoss and Cohen [2013]) or there may not be available beds in the primary unit requested. The first type of patients typically require transfers from critical units to intermediary care units and we will refer to them as "critical transfers", whereas other patients will be referred to as "overflow transfers" from now on.

From a medical point of view, overflow transfers hinder the quality of care, thus they are not desirable (West [2010a], Association [2013]). Also these hand-offs between units lead to discontinuity in the care of patients (Cohen and Hilligoss [2010]). However, from a practical perspective, transfers are almost unavoidable. It is a common strategy for hospitals to cope with bed congestions (Shi, P. and Chou, M. C.

and Dai,J.G. and Ding, D. and Sim, J. [2012]). Thus, incorporating this into our simulation framework is essential for developing a more accurate patient flow model.

The "critical transfers" are already incorporated into the simulation model, with the transfers from critical care units, ED and PACU, using average values specific for each MDC category. We have initially failed to incorporate the overflow patients into our simulation framework, because the data-set from BMC does not include patient transfers. As an extension, we added an algorithm to integrate interdepartmental transfers by allowing overflow transfers for units with the highest interactions. In order to so, we establish a set of alternative units for each admit unit as a result of our data analysis. A patient is overflown to a secondary unit, if the queue size for the primary unit is greater than a pre-specified threshold, and the alternative unit's queue size is less than the primary unit's queue.

The preliminary analysis of the outputs show that incorporating overflow transfers has a significant impact on decreasing waiting times and queue size (see Appendix E for the results). In order to have more conclusive results, a simulation that incorporates transfers should be analyzed for 10 replications. Also the relationship between transfers and the discharge policies can be investigated to investigate questions like how many discharges should happen in a congested unit to avoid the overflow transfers. This analysis will enable us to construct a trade-off curve between number of discharges and transfers.

A different way to model transfers is to incorporate an overflow policy. For instance, in the hospital Shi, P. and Chou, M. C. and Dai,J.G. and Ding, D. and Sim, J. [2012] is working with, National University Hospital (NUH), there exists

a guideline on when and how to overflow a patient. The hospital would overflow patients more aggressively during late night and early morning (before 7 AM). The reason is that few discharges happen in this time period, so there is little chance that a bed in the primary unit will become available in the next few hours. After understanding the overflow policy of BMC, using the expert idea of our nursing collaborators, a policy similar to this can be incorporated.

4.11 Hospitalist Scheduling Problem

After briefly summarizing some of the literature on hospitalist operations, I will provide an overview of the hospitalist scheduling problem, followed by some of the field observations from BMC.

4.11.1 Literature review

The term "hospitalist" was first used in 1996 (Wachter and Goldman [1996]). Simply put, physicians whose primary professional focus is hospital medicine are called hospitalists. Hospitalists, most of whom are board-certified internists (internal medicine physicians), coordinate the care of patients in the hospital. They are the equivalent of PCPs in the inpatient setting and are vital to the flow process, since hospitalists are familiar with the hospital's system and maintain close relationships with specialty physicians.

There are three main types of hospitalists: 1) Admission hospitalists who are solely responsible for admitting patients. 2) Rounding hospitalists who are responsible for the discharge and the overall care of the patients. There are 2 types of

rounding hospitalists: a) Teaching (Attending physicians mentoring residents and medical students), b) Non-teaching (Only attending physicians).

Six years after their first article, Wachter and Goldman [2002] published another paper supporting the premise that hospitalists improve inpatient efficiency without harmful effects on quality or patient satisfaction, justified by empirical evidence. Even though, there is supporting evidence that hospitalists improve efficiency, there is little data that explains how hospitalists achieve this. Tipping et al. [2010] find that only 17% of a hospitalists' time is spent as direct patient care and 64% for indirect patient care, which is mostly spent on working with electronic medical records (EMR). These EMR activities mostly include progress notes and discharge instructions. Travel time accounts to 6% of their time. Thus interestingly the authors find that hospitalists spend more time reviewing and updating the EMR than directly with the patient.

Other studies with similar findings are published by a couple of groups: O'Leary et al. [2006], Maguire [2010] and Kim et al. [2010] find that hospitalists on average spent 18% of their time on direct patient care, 69% on indirect patient care, 4% on personal activities, and rest on education and travel. From the 69% for indirect patient care, communication accounted for 24% of their activities, with 6% used for paging other physicians, and 7% for returning pages. The hospitalists at the study believed they frequently performed simultaneous activities and were excessively interrupted by pagers. As a solution, the authors discuss that 2 way pagers could facilitate communication to decrease the unnecessary interruptions (O'Leary et al. [2006]).

One problematic area in hospitalist schedules is "batching" behavior. Maguire [2010] focus on the impact of batching, which is described as "cyclicality of hospitalist workflow". This leads to problems for other departments, for instance labs, radiology and pharmacy and congestions throughout the hospital. Kim et al. [2010] also discuss the batching behavior of the hospitalists, which cause delays and spikes in indirect care followed by spikes in direct patient care, especially at shift changes.

Even though the workload of a hospitalist varies significantly, typically hospitals target a workload of 12-15 patients per day per physician, depending on whether the physician has support from residents or not. A lower number applies when physicians are working without house staff while a higher number is used when hospitalists are serving as attendings on the teaching service. According to Colwell [2013], if you ask a group of hospitalists how many patients they can manage without feeling overburdened, answers would range from 10 to 20 or more. Nationally, hospitalists care for an average of 11.3 patients per shift.

Colwell [2013] also discusses that "there is a point beyond which patient care suffers, when hospitalists are not able to make good decisions and keep track of everything". Hospitalists in academic institutions experience high levels of burnout and have relatively little opportunity for scholarly work (Dyrbye and Shanafelt [2011]). In a survey that was performed in big medical centers, 67% of hospitalists reported high levels of stress, and 23% being burnout (Kling [2011]). More importantly, 40% of the hospitalists nationwide reported that inpatient census levels exceeded "safe" levels at least once a month, and 36% having an unsafe workload at least once a week. This leads to physicians ordering unnecessary tests, procedures, or consultations as well

as not being able to answer questions from the patients due to inadequate time with the patient and eventually leads to decline in care and patient satisfaction (Glasheen et al. [2011]). As the life expectancy increases, patients grow older, and the difficult cases with multiple chronic conditions increase, hospitalists will have to deal with more complex case management and thus will face a greater workload. Thus, creating a better schedule for hospitalists is also vital to the delivery of high-quality health care.

4.11.2 The problem

One of the main contributing factors to long waiting times is the delayed discharges. Typically, hospitalists are responsible for patient admissions, monitoring and discharge process. Hospitalists experience high levels of stress and burnout, because of their high workload. On their morning rounds hospitalists need to both care for patients admitted overnight, acute inpatients and also discharge the patients who are completing their LOS. The timing of discharges is closely related to how hospitalists prioritize patients on their rounds. Priority overall is typically given to more severe patients over discharges, because newly admitted patients have more urgent needs. This leads to delayed discharges, which results in delays for patients who are waiting for an inpatient bed.

Physical layout of the hospital further complicates the scheduling problem, and introduces a geographic, traveling salesman type component to the model. To the best of our knowledge, no study has tackled this problem. Modeling the patient mix while optimizing the hospitalist rounds in order to improve the patient flow will be

a significant extension to our current research on inpatient capacity management. It will be crucial to link the problem of hospitalist scheduling with the results of simulation model that suggests effective discharge policies.

The optimization/simulation framework will have a multiple objective structure while incorporating the impact of case-mix on the workload of a hospitalist. Potential objectives include:

- Minimizing the number of hand-offs.
- Minimizing traveling time.
- Maximizing patient diversity.
- Maximizing the number of patients seen without an adverse impact on quality of care.
- Maximizing the number of early in the day discharges for units that are full.

The main decision variable is the number of patients assigned to each hospitalist from each admit unit every hour. In other words, what is the optimal panel size and case-mix for a hospitalist? How does a hospital allocate the limited discharge capacity to units? It will be an essential step to integrate the results of our simulation, which finds effective discharge policies, with the schedules of the hospitalists.

4.11.3 Observations from BMC

I shadowed three different types of hospitalists for 3 days and our observations are similar to the literature. In BMC, these hospitalists work under three shifts: Regular

hours (8 AM-6 PM), evening (3 PM-9 PM) and night (8 PM-8 AM) shift. The majority of the hospitalists are employed during the regular hours. Case managers also work with hospitalists on the discharge process. They are typically nurses and are responsible for arranging the post-acute care of patients like nursing and rehab facilities.

Admission hospitalists are solely responsible from admitting the patients, ordering initial labs, contacting the PCPs and performing reconciliation (reviewing the list of medications the patient is taking and updating the dosage during the hospitalization period). They are not responsible for the follow-up of the patients after the admitting decision is made, rather the rounding hospitalists are responsible for their care in the hospital.

For the rounding hospitalists on teaching service, there is no geographical alignment as the diversity of patients is crucial for residents' training. On the other hand, the non-teaching groups work based on geographical alignment so less time is spent traveling.

The day of a hospitalist can be summarized in 5 main groups: direct patient care, indirect patient care, travel, education, professional development. The direct patient care involves taking the history, initial examination, meetings with the family members, seeing patients during the follow-up visits and providing discharge instructions. Indirect patient care is composed of communication, documentation, writing orders, initiating and returning pages, reviewing test results and medical records. The education component involves teaching during the rounds.

In our limited time-motion study we observed that hospitalists spent most of their time on indirect patient care activities and relatively little time on direct patient care. Table 4.7 demonstrates some of the observations from our time-motion study, supporting that indirect patient care constitutes the majority of the hospitalists' workload. Studies in the literature have reported similar findings (O'Leary et al. [2006], Tipping et al. [2010]).

Table 4.7: Time motion study

Patient #	**Examination**	**Indirect patient care**	**Travel**
1	2	10	1.5
2	6	4	0.5
3	6	8	1
4	1.5	6	2.5
5	13	3	3
6	1.5	3	1.5
7	3.5	4	0
8	2	1.5	1.5
9	1	3.5	1.5
10	2	4	3
11	1	5	1.5
12	5	3	0
13	1	3.5	3
14	2	6	3
Average	3.4	4.6	1.7
Std dev	3.2	3.2	1.0

CHAPTER 5

OPTIMIZING SPINE OR SURGICAL THROUGHPUT: ENGINEERING A PULL SYSTEM FOR OUTPATIENT ACCESS

5.1 Introduction

For spine surgeries, large medical centers like Mayo Clinic generally face more patient demand than available capacity. One reason is the relatively long surgical times for spine patients. Data from Mayo Clinic shows that 50% of spine surgeries are over 4 hours in length. Thus, on most days a spine surgeon is able to do only one or at most two surgeries. This limits patient access and may cause significant delays for surgery scheduling.

Due to the length and variability of spine surgeries scheduling is a difficult and important aspect to patient access, effective operations, and financial performance for the spine surgery practice (Dexter et al. [2010]). In addition, as noted in Espin et al.

[2006] safety for both the patient and surgical staff may be an issue if surgical days run long. Further complicating scheduling and operating room (OR) management is emergency cases, short-term cancellations, and complex cases that require more than one surgery to address a patient's needs.

These factors together create significant uncertainty for scheduling spine surgeries. At Mayo Clinic this resulted in 38% of surgical days going past the desired end time of 5 PM. Overtime is a significant issue at Mayo Clinic due to the importance of quality of life for the surgeons and the surgical teams in addition to the aforementioned concerns regarding safety. At the same time, OR utilization during normal hours was less than desired, limiting patient access and reducing potential financial performance.

5.1.1 Historical spine surgery scheduling at Mayo Clinic

Many of the concepts and approaches discussed in this research are relevant to other surgical practices and particularly those in spine surgery, nonetheless, the orthopedic spine surgery practice at Mayo Clinic has many unique characteristics. In this section we discuss the specific problem setting.

Mayo Clinic's core value is the "needs of the patients come first." This influences surgical scheduling because patient timing needs (preferences) are important to final scheduling decisions. This is in contrast to many surgical settings where patients are simply told when to show up. Thus, at Mayo Clinic the patient discusses with the surgeon and their team, when to schedule their surgery. Due to the fact that spine surgeries often have significant impact on patients' lives for extended periods and

the lengthy recovery process, the patients' scheduling preferences are important to consider.

However, this approach led to problems in daily surgical loads. If a patient preferred their surgery on a particular day or week where several other surgeries were already scheduled it may have led to significant overtime. In part, this was due to the difficulty in simply "squeezing in" another spine surgery due to their length and variability. Conversely, other days and weeks were underutilized. In the absence of good information regarding the current status of their schedule, surgeons and those doing their scheduling were often driven to make decisions influenced too much by patient preferences.

Scheduling surgeries at Mayo Clinic is further complicated by the fact that dedicated OR time is available to most surgeons. This is a positive in that it allows the surgeons a great deal of autonomy in managing their cases. However, it is problematic in that the organization cannot pool OR time and balance loads across all ORs. Rather each surgeon's load must be balanced across their surgical days. These surgical days are assigned via the "Blue and Orange" system at Mayo Clinic. This harkens back to the system developed by the Mayo brothers, Dr. Will and Charlie, who performed surgeries every other day, in complementary fashion. In the first week, one surgeon ("blue") performs surgeries on Mondays, Wednesdays and Fridays, while the other surgeon ("orange") is active on Tuesdays and Thursdays. In the next week, the orange surgeon performs surgeries on Monday, Wednesday and Friday while the blue surgeon does so on Tuesday and Thursday. This alternating cycle is then repeated. On days that surgeries are not performed, the surgeon sched-

ules clinical consultations with patients. This created a simple management system for clinic and surgery days that continues today, but results in some restrictions in scheduling flexibility and can make short-term case load imbalances worse as some surgeons get overloaded and others underutilized.

While Mayo Clinic is a non-profit organization, financial viability and sustainability is still an important consideration. Profits from clinical practice support research, education, and ongoing improvement initiatives, all of which are important to Mayo Clinic's mission. With limited capacity to allocate to the high demand for spine surgeries, some control of which surgeries are performed and when, can be important to net operating income (NOI). NOI is a measure of the projected revenue less operating and fixed allocated costs. Given specific revenue reimbursements and Mayo's cost structure, some types of spine surgeries are more profitable than others. Note that patients with government payers (e.g., Medicare and Medicaid) generally have lower profitability, but there was no desire to reduce the number of such patients.

Further the overall patient profitability to Mayo Clinic, including their hospital stay is affected by the timing of surgeries. A significant proportion of spine surgery patients require discharge to a skilled nursing facility (SNF). These facilities generally do not accept patients on weekends and therefore if a patient requires a SNF and their planned discharge is on a weekend, Mayo often incurs the additional costs without compensating revenue. This is because government insurance payers generally have a fixed reimbursement for each procedure type. It may be difficult to a priori determine the risk of a patient requiring a SNF at discharge, however, it is known that older patients have a higher risk and because such patients were generally covered by

Medicare, special attention is paid to when these patients were scheduled. In general, these patients were scheduled on Mondays and Fridays with the assumption that this would result in the least number of delayed discharges. It is important to recognize that while all the above factors are important when scheduling patients, the Mayo system needed to have the flexibility to ensure that the needs of the patients always come first.

5.1.2 Objectives

The primary objective of our research is to create better patient access as a result of increased surgical capacity with more efficient schedules. We not only maximize surgeon and OR utilization but also incorporate profitability while keeping overtime and potentially unsafe surgical days under control. In addition, the proportion of government payer patients was set to at least be maintained at historical levels. Because the overall objective was to increase patient access, this constraint should actually increase the number of Medicare and Medicaid patients treated at Mayo Clinic.

5.1.3 Approach

Our approach involves seven steps: First, we perform data analysis to identify categories of surgeries, that can be grouped together based on the their surgical durations. Next using these surgical categories, a simulation model is used to identify feasible surgical pairs that can be performed in a day. The surgical pairs and their outcomes are then used in the first stage optimization model to maximize a weighted combination of utilization and net operating income. This generates the optimal surgical

case-mix to be performed from each surgery category on each day over a planning horizon. The second stage optimization model creates the optimal schedule using the results of the first stage model and maximizes available slots for complex multiple days staged surgery cases. The last stage of the optimization creates a schedule that remains feasible to the requirements of the hospital, by incorporating Mayo Clinic specific scheduling requirements (blue-orange surgical template). A second simulation model is developed and used to test the impact of urgent surgeries and cancellations on the optimal schedule. As a last step, our optimization framework was implemented in Mayo Clinic in a controlled pilot and we evaluate the results of the intervention.

5.2 Literature Review

Surgery scheduling has three main decision levels: strategic, tactical and operational level. These levels represent long, medium and short term decisions respectively. These typically refer to case-mix planning, master surgery scheduling and case scheduling. Case-mix planning assigns available OR time to specialties, whereas master surgical scheduling creates a recurring cyclic timetable (Guerriero and Guido [2011]).

Some of the most commonly analyzed problems, as identified by Gupta [2007], include sequencing surgical cases, allocating elective surgical capacity to different sub-specialties, creating a booking limit for elective cases, finding the optimal sub-specialty mix and creating a master surgical schedule (MSS) with a rolling horizon.

There is a wide variety of papers on surgery scheduling in the literature. And various approaches have been used to optimally schedule surgeries to ORs like, integer programming (Blake and Donald [2002], Denton et al. [2007], Denton et al. [2010], Vissers et al. [2005], Adan et al. [2009]), stochastic optimization models (Denton et al. [2010], Batun et al. [2011], Van Oostrum et al. [2008], Testi et al. [2007]), goal programming (Rohleder et al. [2005]), discrete event simulation (Adan et al. [2009]), and heuristics (Denton et al. [2010], Van Oostrum et al. [2008], Testi et al. [2007]).

Rohleder et al. [2005] use a goal programming approach to schedule surgery blocks to an OR schedule with the objective of smoothing post-surgery patient flow. Their formulation is similar to Blake and Donald [2002] who use an integer programming approach to create a master surgical schedule. Both of these papers and most of the formulations in the literature use deterministic models. Whereas, we take into account the stochastic nature of surgery durations and use scenarios (derived from the simulation model) in our integer program. There are a limited number of papers that model stochastic surgical durations (Denton et al. [2010], Batun et al. [2011]).

The objectives in the formulations range from smoothing post-surgery patient flow (Denton et al. [2010]), minimizing the over and under-utilization of multiple resources (Van Oostrum et al. [2008], Testi et al. [2007]), minimizing the weighted sum of the expectation of waiting time, idle time, and tardiness (Denton et al. [2007]), minimizing the deviation from the target utilization level (Vissers et al. [2005]), to minimizing both the overtime and fixed cost related with opening an OR (Denton et al. [2010]).

Accordingly, the decision variables of the models vary as well. Most commonly, authors look at block assignment decisions and development of a cyclical master schedules which involves assigning the number of sessions to each sub-specialty (Adan et al. [2009], Van Oostrum et al. [2008], Testi et al. [2007]). On a more daily level, the authors look at decisions for sequencing surgeries in ORs, which ORs to use, and the start times of each surgery (Batun et al. [2011]). Some papers assume that the type of surgeries to be performed in a day is predetermined and only focus on the sequencing of surgeries in an OR (Choi and Wilhelm [2012]). The common result is that the smallest variance first (SV) rule gives the best sequencing decision, and that the decision on the first surgery is the most crucial (Weiss [1990], Van Oostrum et al. [2008], Denton et al. [2007]). In our model, we do not consider the sequencing of the surgeries.

5.2.1 Contributions

The problem we address is motivated by a specific case study at Mayo Clinic and is unique compared to the previous surgical scheduling research. Due to the long and highly variable nature of spine surgeries, only one or at most 2-3 surgeries can be performed within the ten hours of operating time available at Mayo Clinic. Therefore, sequencing does not play an important role. However, because of the high variability of surgery times, maintaining both high OR utilization and low overtime is challenging. To assist the practice, our model considers alterations to the patient mix and identifies which surgeries can be performed to achieve acceptable overtime

levels. The patient categorization using clinical information known at the time of case scheduling is also unique and can be applicable to other surgical areas.

In addition, we consider constraints on financial performance and at the same time ensure access to Medicare patients. To the best of our knowledge, we are one of the first to consider the impact of patient mix, surgical schedule and LOS on the financial performance. In order to achieve this, we identified the best days to perform each type of surgery based on the hospitalization period, to avoid uncompensated weekend stays.

We consider a multiple objective, multiple surgeon/OR surgical case assignment problem with stochastic surgical durations. However, it is not a block scheduling problem rather, we are creating a cyclical surgical schedule specific for the spine surgery clinic that assigns different types of surgical cases to days of the week while optimizing the surgical case-mix.

Lastly, the literature on surgical scheduling is quite extensive, however, implementations are rare (Cardoen et al. [2010]). Namely, Blake and Donald [2002] is one of the rare applications that develop a deterministic model without considering the variability in surgical durations. The authors do not provide details on the process of implementation. Indeed, there is lack of information on the behavioral factors that influence the actual implementation and identification of the causes of failure or the reasons that lead to success (Cardoen et al. [2010]). We believe that our main contribution is being able to evaluate the results of our pilot study with a pre and post evaluation, as well as using a control and a test group. Secondly, despite the wide variety of literature on OR scheduling problems, the multi-OR surgical suite

scheduling problem and the impact of addressing demand uncertainty have not been studied thoroughly (Erdogan et al. [2011]). With this research our objective is to contribute to both of these fields.

5.3 Optimization Models

As described in Section 5.1.3, our optimization model involves multiple stages; if all stages are considered together, tractability becomes an issue. The stages also represent the corresponding decision level as we move from a strategic (first stage) to a daily decision level (third stage). The first stage is a strategic level decision that decides on the optimal patient case-mix in a given time horizon in order to maximize a weighted function of utilization and estimated profitability (via NOI). With this optimal surgery mix as input, the operational decision level (second stage) allocates cases to specific days in the time horizon, while ensuring that multiple days staged surgeries (performed on the same patient) can be carried out within a few days. The third stage assigns the surgeries to operating rooms, using the surgical template from second stage, while balancing the workload between surgeons and incorporating Mayo's blue-orange surgery template.

Before describing the 3 stages, we note that there are 3 sets of indices for days in the formulation; k, d and t. k represents the index for group of day which can either be a late start day or a regular start day (late starts happen due to teaching responsibilities on specific days, and the start time influences overtime measures); d represents the day of week (Monday to Friday, which help in differentiating hospital

specific dynamics such as blue and orange surgical days described in Section 5.1.1); and t is the index for days in the time horizon.

5.3.1 First stage IP optimization

In this stage we find the optimal surgery mix whilst maximizing a weighted combination normalized NOI and utilization with Equation 5.1. This stage uses the outputs of the simulation (NOI, utilization and overtime percentage derived for each surgical combination) as an input to the optimization model. The formulation is as follows:

Indices

$i(1...I)$: Combination of surgeries

Each combination i consists of some surgery category l and a payer type r associated with it.

$k(1...K)$: OR-weekday category (where 1 means a regular weekday; and 2 is a late start day)

$d(1...5)$: Day of week (Monday, Tuesday, Wednesday, Thursday and Friday)

$l(1...L)$: Surgery category index

$r(1..R)$: Payer index (where 1 is for Medicare or government; and 2 is for non-Medicare or private)

$t(1...T)$: Days in the planning horizon

Parameters

OT_{ik}: Simulation derived parameter representing the probability of finishing after the end of day (5 PM) when surgery combination i is performed on day group k. Re-

stricted to be less than "o" in the optimization model which is the practice imposed limit on the proportion of overtime after 5 PM.

EO_{ik}: Simulation derived parameter representing the expected overtime observed after end of day (5 PM) when the surgery combination i is performed on day group k. Restricted to be less than "e" in the optimization model which is the historical limit on the expected overtime after 5 PM.

θ_{ik}: Simulation derived parameter representing the probability of finishing after 11 PM when surgery combination i is performed on day group k. Restricted to be less than "f" in the optimization model which is the empirical limit on the percentage of overtime after 11 PM.

NOI_i: Simulation derived parameter for normalized NOI of surgery combination i.

U_{ik}: Simulation derived parameter for the average OR utilization when surgery combination i is performed on k^{th} day group.

ω: Weight assigned to utilization in the objective function.

M_{ilr}: The number of category l surgeries in each surgical combination i with payer r.

T: Number of working days in the planning horizon.

P_l : Empirically observed number of surgeries from surgical category l per OR room.

b: The case-mix bound width represented as a fraction between 0 and 1 (i.e. allowed flexibility in changing the case-mix).

m: Minimum percentage of Medicare surgeries to be performed.

F_{ld}: Binary parameter that takes on the value of 1 if day of week d is the best surgical day for category l Medicare patients; 0 otherwise.

$\sum_d F_{ld} = 1$ and F_{ld} binary $\forall l, d$.

D_{kd}: Binary parameter that takes on the value of 1 if day of week d is a type k day (i.e. if it is a regular start or late start day); 0 otherwise.

$\sum_k D_{kd} = 1 \ \forall d$ and D_{kd} binary $\forall k, d$.

J_t: The open number of ORs on day t.

B: Large integer constant.

Decision Variables

x_{ik}: Total number of surgery combinations of type i performed on day group k over the time horizon T.

ζ_{ld}: Output variable representing the number of Medicare surgeries from surgical category l scheduled on day of week d.

First Stage Model

$$\max \ \omega * \sum_k \sum_i U_{ik} * x_{ik} + (1-\omega) * \sum_k \sum_i NOI_i * x_{ik} \tag{5.1}$$

s.t

$$P_l * (1-b) * J_t \leq \sum_k \sum_i \sum_r M_{ilr} * x_{ik} \leq P_l * (1+b) * J_t \ \forall l \tag{5.2}$$

$$\sum_k \sum_i x_{ik} * M_{il1} \geq \sum_i \sum_k \sum_r x_{ik} * M_{ilr} * m \tag{5.3}$$

$$P_l * (1-b) * J_t * m \leq \sum_k \sum_i x_{ik} * M_{il1} \leq P_l * (1+b) * J_t * m \ \forall l \tag{5.4}$$

$$\sum_i \sum_k \frac{OT_{ik} * x_{ik}}{\sum_t J_t} \leq o \tag{5.5}$$

$$\sum_i \sum_k \frac{EO_{ik} * x_{ik}}{\sum_t J_t} \leq e \tag{5.6}$$

$$\sum_i \sum_k \frac{\theta_{ik} * x_{ik}}{\sum_t J_t} \leq f \tag{5.7}$$

$$\sum_i \sum_k D_{kd} * x_{ik} = \sum_w J_{5*w+d} \quad \forall d \tag{5.8}$$

$$\zeta_{ld} = \sum_i \sum_k x_{ik} * M_{il1} * F_{ld} * D_{kd} \; \forall l, d \tag{5.9}$$

$$\sum_i \sum_k x_{ik} * D_{kd} * M_{il1} \leq B * F_{ld} \; \forall l, d \tag{5.10}$$

$$x_{ik} \in \mathbb{Z}_{\geq 0} \; \forall i, k \tag{5.11}$$

There are four main groups of constraints in this stage: case-mix and payer mix calculations (Equations 5.2-5.4), overtime restrictions (Equations 5.5-5.7), restricting the number of surgeries based on the open number of ORs (Equation 5.8), and enforcing that Medicare surgeries are scheduled on best day of the week to minimize the number of weekend discharges (Equations 5.9-5.10). We explain these in more detail below.

In order to build a realistic model, the surgical schedule needs to create a surgical case-mix that is similar to observed levels in the current practice. Thus, Equation 5.2 ensures that number of patients from surgical categories ($1...L$) are only allowed to deviate from the current case-mix within a pre-specified bound width, b. Medicare surgeries constitute at least $m\%$ of the overall number of surgeries performed with Equation 5.3. Further, Equation 5.4 enforces that the sum of Medicare patients from each surgical category is within the $\pm m\%$ range of the empirically observed number of patients.

The percentage of days that result in overtime (percentage of days that end after 5 PM) is enforced to be smaller than some overtime limit based on the clinic's preference, "o" (Equation 5.5). Expected hours of overtime after 5 PM is kept less than the empirical average overtime hours, "e" (Equation 5.7). Similarly, the proportion of days with overtime after 11 PM is required to be less than the empirical overtime limit based on historical data, "f" (Equation 5.6).

Equation 5.8 ensures that the total number of surgeries that will be performed on each day of week d in the horizon must be equal to the total number of operating rooms available on such weekdays in the horizon.

In Equation 5.9, ζ_{ld} represents the number of Medicare surgeries from category l scheduled on day of week d; which is simply the product of a binary variable that indicates whether or not that day was indeed the best day to do the surgery for a category l Medicare patient day and the sum of all Medicare surgeries for that specific surgery category. With Equation 5.10, Medicare patients are assigned to their best day of surgery, specific for each category so that the Medicare weekend overflow is minimized. This has a huge financial impact, which will be discussed later in the case study using empirical data.

5.3.2 Second stage IP optimization

Using the case-mix results of the first stage optimization as an input, we create a schedule over the planning horizon, by assigning the surgery combinations to days in the horizon. The surgical schedule repeats itself every T days.

The main objective in the second stage is to maximize the availability of ORs for complex surgeries performed on the same patient staged over multiple days. We call them "multiple days staged surgeries" (MDSS). These result when some of the very long surgeries are broken down into 2 or more surgeries with feasible durations by the surgeons; that need to be carried out within 2-5 days. Data analysis indicated that these types of surgeries constitute a non-negligible 10% of the overall surgeries performed. For example a patient may need to undergo a category 6 surgery followed by a category 8 surgery within two days (this would be MDSS type 6_8). If t is the day of the first surgery, then this means that a combination containing surgery 6 must be scheduled on day t and a combination containing surgery 8 on day $t+2$. One of the ways this would be possible is if combination 1_6 is scheduled on day t and combination 1_8 is scheduled on day $t+2$. To ensure MDSS constraints are met in the formulation, we use a binary parameter δ_{ics} that takes on the value of 1 if surgery combination $i = (1_6)$ contains one element of the MDSS of type $s = (6_8)$ in cth position; for this example if $c = 1, \delta_{ics} = 0$, but if $c = 2, \delta_{ics} = 1$.

Note that, this stage has no impact on NOI or utilization, since they have already been optimized in the first stage.

Indices

$i, j(1...I, 1...J)$: Combination of surgeries

$s(1...S)$: MDSS index

$c(1...C)$: Position in the sequence in which the MDSS is performed

$w(1...W)$: Weeks in the planning horizon

Parameters

ζ_{ld}: Number of Medicare surgeries from surgery category l scheduled on day of week d (Derived from the first stage).

β_s: Weight of type s MDSS in the objective function.

α: Coefficient for balancing the workload over the weekdays.

δ_{ics}: Binary parameter that takes on the value 1, if surgery combination i contains one element of the MDDS s in cth position; 0 otherwise.

Decision Variables

Y_{it}: Integer decision variable representing how many of the surgery combination i's are performed on day t as a part of multi-surgery pair.

Z_{it}: Integer decision variable denoting how many of the surgery combination i's are performed on day t not as a part of MDSS (rather as a single stand-alone surgery combination).

L_{ts}: Integer decision variable denoting how many of the surgeries on day t are performed as the first component of the MDSS type s.

Q_{id}: Number of surgery combination i's to be performed on day of week d.

Second Stage Model

$$\max \sum_{t} \sum_{s} (\beta_s * L_{ts}) \tag{5.12}$$

s.t

$$\sum_{d} Q_{id} D_{kd} = x_{ik} \ \forall i, k \tag{5.13}$$

$$\sum_{w} (Y_{i(5w+d)} + Z_{i(5w+d)}) = Q_{id} \ \forall i, d \tag{5.14}$$

$$\sum_{i} \sum_{w} Y_{i(5w+d)} M_{il1} + \sum_{j} \sum_{w} Z_{j(5w+d)} M_{jl1} = \zeta_{ld} \ \forall l, d \tag{5.15}$$

$$\sum_{i} Y_{it} + \sum_{j} Z_{jt} = J_t \ \forall t \tag{5.16}$$

$$L_{ts} \leq \sum_{i} \delta_{i1s} * Y_{it} + \sum_{j} \delta_{j2s} * Y_{j(t+2)} \ \forall t, s \tag{5.17}$$

$$L_{ts} \leq \sum_{i} \delta_{i1s} * Y_{it} \ \forall t, s \tag{5.18}$$

$$L_{ts} \leq \sum_{j} \delta_{j2s} * Y_{j(t+2)} \ \forall t, s \tag{5.19}$$

$$\sum_{w} \sum_{s} L_{(5w+d)s} \geq 5\alpha \sum_{t} \sum_{s} L_{ts} \ \forall d \tag{5.20}$$

$$Y_{it}, Z_{it} \in \mathbb{Z}_{\geq 0} \ \forall i, t \tag{5.21}$$

$$L_{ts} \in \mathbb{Z}_{\geq 0} \ \forall t, s \tag{5.22}$$

$$Q_{id} \in \mathbb{Z}_{\geq 0} \ \forall i, d \tag{5.23}$$

The objective function maximizes the weighted sum of MDDS performed (Equation 5.12). The weights (β_s) are derived from the empirically observed proportions of type s MDSS. Detailed analysis is presented in Section 5.4.1.2.

Constraints

Equation 5.13 links the first stage output (the optimal surgery case-mix), to the second stage decision variable (number of surgeries scheduled from each surgery category on specific days of week). Equation 5.14 ensures that number of combination

i surgeries performed on each day of week matches with the first stage results, via the Q_{id} decision variable. Next, Equation 5.15 enforces that the required number of Medicare surgeries are performed on the right day of week for each category.

All in all there can at most be J_t number of combinations scheduled each day (Equation 5.16); recall that J_t is the number of open/available operating rooms on day t. Equations 5.17, 5.18 and 5.19 ensure that for a MDSS of type s to take place, the second surgery of MDSS pair needs to be arranged within 2 working days after the first surgery. With Equation 5.20 MDSSs are spread evenly over the workdays, using a lower bound α. This ensures that not all of the MDSSs are performed on the same days of week.

5.3.3 Third stage IP optimization

The last stage of the optimization model uses the surgery template generated from second stage to balance the surgeons' workloads over the days of the week, while incorporating Mayo specific scheduling requirements. In Mayo Clinic, the surgeons operate under the blue-orange schedule, in which they perform surgery on one day and have clinical consultations on the next day, as discussed in Section 5.1.1. For pairs of surgeons, these surgical and consultation days alternate.

Indices

$h(1...H)$: Surgeon indices

$q(1...Q)$: Types of surgical weeks

Parameters

λ_{hdq}: Binary parameter matrix denoting if day of week d on week type q is the surgery day for surgeon h.

For instance, if we look at this matrix for Surgeon 1 (a blue surgeon) under a blue-orange surgical schedule (where $Q = 2$), λ_{1dq}, would be equal to: $\begin{bmatrix} 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \end{bmatrix}$. This implies that Surgeon 1 would be operating on Monday, Wednesday and Friday the first week type ($q = 1$) and on Tuesday and Thursday on the second week type ($q = 2$).

Decision variables

ϕ_{qw}: Binary decision variable denoting if the week w is of week type q. For example, $\phi_{q'w'} = 1$ implies that week w' is of type q'.

τ_{lhr}: Absolute difference of workload for surgeon h, from the average number of category l surgeries scheduled over the planning horizon with payer r.

W_{lhr}: Number of category l surgeries with payer r scheduled over all weeks for surgeon h.

Third Stage Model

$$\min \sum_{l} \sum_{h} \sum_{r} \tau_{lhr} \tag{5.24}$$

s.t

$$\sum_{w} \sum_{q} \phi_{qw} = W \tag{5.25}$$

$$W_{lhr} = \sum_q \sum_w \sum_d (\sum_i (Y_{i(5w+d)} + Z_{i(5w+d)}) M_{ilr} \phi_{qw} \lambda_{hdq} \ \forall l, h, r \tag{5.26}$$

$$\tau_{lhr} \geq W_{lhr} - \frac{\sum_{h'=1}^{H} W_{lh'r}}{H} \ \forall l, h, r \tag{5.27}$$

$$\tau_{lhr} \geq \frac{\sum_{h'=1}^{H} W_{lh'r}}{H} - W_{lhr} \ \forall l, h, r \tag{5.28}$$

$$\sum_q \phi_{qw} = 1 \ \forall w \tag{5.29}$$

$$\phi_{qw} \in 0, 1 \ \forall q, w \tag{5.30}$$

$$\tau_{lhr} \geq 0 \ \forall l, h, r \tag{5.31}$$

$$W_{lhr} \in \mathbb{Z}_{\geq 0} \ \forall l, h, r \tag{5.32}$$

The objective is to balance the workload between the surgeons so that this absolute difference is minimized. Index h is the index for surgeons and q represents the different patterns of weeks ($1...Q$). In our case study, we implement the blue orange surgical schedule, however, the model is kept general to accommodate different scheduling patterns in other hospitals.

Constraints

The decision variable ϕ_{qw} is a binary variable denoting if the week w is of week type j. The sum of all types of weeks should add up to W, which is enforced by Equation 5.25.

Each surgeon's workload over the planning horizon is calculated using Equation 5.26. τ_{lhr} is calculated as the absolute difference between the workload of each surgeon and the average number of surgeries from each surgical category, in a linear fashion with Equations 5.27 and 5.28. Lastly, with Equations 5.29 and 5.30 we ensure

that the weeks are of some type q. In Section 5.4.3.3 we provide an example from our case study with 2 surgeons.

5.4 Case Study

The optimization model was developed and evaluated based on the operations and data of the orthopedic spine practice at Mayo Clinic, Rochester MN. Model development and evaluation occurred over much of 2012 with a live implementation target set as December 2012. We have set the planning horizon T to 120 days (with a surgical day length of 10 hours) this is a large enough horizon to observe demand for all the surgery categories, including those that are sparsely represented. We consider two types of surgical days: regular and late start days. The latter occurs on Mondays to allow for staff meetings and reduce the day length by one hour. As discussed in Section 5.1.1, Mayo Clinic surgeons have alternating surgical and non-surgical days based on the "Blue and Orange" system. The non-surgical days are typically spent in clinic where surgeons have follow-up appointments or see new patients who may require surgery.

5.4.1 Data and model assumptions

Spine surgery related data involves 2 primary OR rooms with 5 surgeons performing more than 2500 surgeries over a 7 year horizon from 2005 to 2011. Data available includes patient-related (age, gender, geographical location, American Society of Anesthesiologists (ASA) scores of patient physical condition before surgery, initial diagnosis (ICD9 code), and length of stay(LOS)), surgery-related (surgeon name, OR

room, long description of the surgery performed provided by the surgeons, surgery durations broken down into OR enter to incision time, incision to closure time, and closure time to OR exit time) and financial information (procedures performed, cost and revenue for each case) at a very detailed level.

Some of the crucial characteristics of the system are: average patient age was 57.6, with 45% female patients. Hospital LOS is on average 5.9 days with a standard deviation of 6.5 days. The average OR enter to Incision time is 1.5 hours with a standard deviation (SD) of 0.4, the average incision to closure time is 4.6 hours with a SD of 2.5 hours. The average closure to OR exit time is 0.5 hours with a standard SD of 0.3 hours. More detail on the patient characteristics is provided in Appendix F.

5.4.1.1 Surgery type:

We classified the whole patient population with 10 surgery categories using Classification and Regression Tree (CART) analysis in JMP (version 9.01, SAS Institute, 2010). This data mining analysis enabled us to more accurately predict how long each surgery will take and therefore better plan the surgery days. Figure 5.1 shows the cumulative distributions for the surgical categories and highlights the difference between the categories. For example, surgeries from category 1 always take less than 4 hours to complete, while on average only 50% of all cases take less than 4 hours to complete.

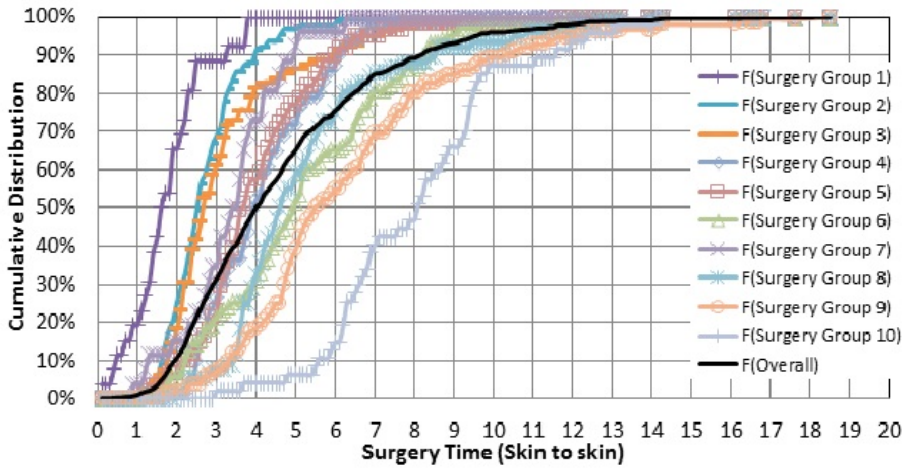


Figure 5.1: Cumulative distribution of the surgery time by each patient category

5.4.1.2 Multiple days staged surgery (MDSS) patients:

Some of the very long surgeries are split into 2 or more procedures with feasible durations. From the data, we determined that these types of surgeries constituted 10% of the all surgeries. These staged surgeries need to be planned within a certain number of days (ideally in 2 days).

Whether a patient will be undergoing a MDSS or not, depends on the American Society of Anesthesiologists (ASA) scores, anatomical location, surgical approach and other factors. We have also analyzed which surgical categories are generally divided into multiple segments, as can be seen in Table 5.1. For instance, a surgery category 6 followed by an 8 in the next 2 days constitute the biggest percentage. These percentages are then used as weights in the objective function of the second stage.

Table 5.1: Properties of multi-segment surgeries

Comb #	Surgery Pair	Proportion (%)
1	6_8	9%
2	8_9	7%
3	4_5	7%
4	8_8	5%
5	1_2	5%
6	7_9	4%
7	4_8	4%
8	4_9	4%
9	2_5	3%
10	3_8	3%
11	5_9	3%
12	1_1	2%
13	1_5	2%
14	1_8	2%
15	2_6	2%
16	2_8	2%
17	4_6	2%
18	7_7	2%
19	7_8	2%
20	1_6	2%

5.4.1.3 Financial analysis and length of stay (LOS):

Our financial analysis is based on the reported Net Operating Income (NOI) values. Data mining approaches were used to derive the cost drivers for the Clinic. NOI values were mainly driven by the LOS of the patients, i.e., the hospitalization period in an inpatient unit post surgery. As well as the LOS, the type of surgery significantly affected NOI. Therefore, depending on the equipments used (such as microscope and CT scanner), characteristics of the surgery (fusion, no fusion, number of vertebrae segments and so forth), the cost of surgeries varied significantly.

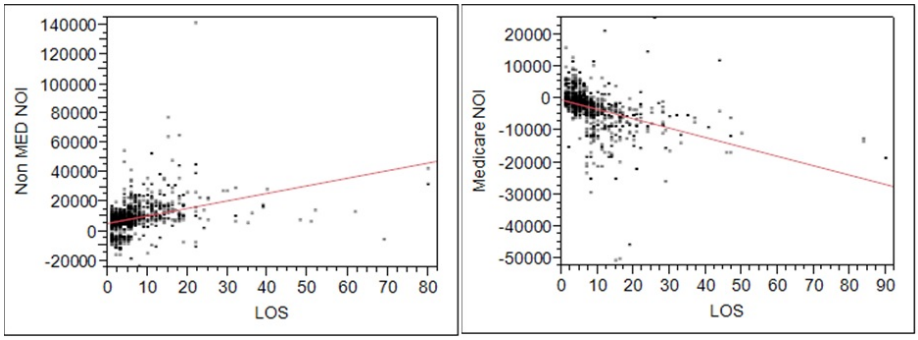


Figure 5.2: Relationship between NOI and LOS for Medicare and Non-Medicare patients

We have performed the NOI analysis for Medicare and Non-Medicare patients separately, because of the reimbursement policies (Figure 5.2). Regardless of their hospitalization period, Medicare patients only get reimbursed for 4 days and the hospital almost always loses money for the Medicare surgeries. However, for non-Medicare patients, hospital is reimbursed depending on the number of hospitalization days. Thus, as can be seen from the graph as the LOS days increase for Non-Medicare patients so does the NOI. However, the opposite is true for Medicare patients. Only some of the patients with small LOS indicate a potential gain, the majority of patients result in negative NOI. In performing this analysis, we have only considered first surgeries of the day, so that the effect of overtime is discarded.

Furthermore, the LOS values of Medicare patients are typically higher than non-Medicare patients (The average is a day longer for Medicare patients). This is mostly because the Medicare patients are mostly elderly and it takes a longer time for the elderly patients to recover. Also, another reason is that they typically require post-acute care, leading to delays in the discharge process.

The initial analysis showed that the optimal surgery day for Medicare surgeries required further examination. Because these patients often require discharge to skilled nursing facilities (SNF) that do not accept patients on weekends, it was important to schedule surgeries to avoid unnecessary weekend stays in the hospital. We calculated LOS in base 7 to analyze their discharge day of the week after the surgery. For example, a LOS value of 8 was equal to 1, meaning the discharge happened on the following day of week (DOW) of the surgery. We look at the probability of weekend overflow by DOW of the surgery and surgery category. Figure 5.3 represents the percentage of patients who are ready to be discharged on any weekend if they have their surgery on that DOW. Our analysis has shown that Mondays and Fridays are generally the best days to schedule surgeries, where it is important to avoid unnecessary weekend stays. However, this is not true for all surgery categories. In particular, some of the more complex surgeries were better to schedule in mid-week due the LOS distribution.

We integrate the optimal day of Medicare surgeries into the first part of our optimization model using the F_{ld} parameter, to minimize the weekend overflows in the optimal solution. Using the results of the data analysis, this binary parameter takes on a value of 1 if the best day of week to perform category l Medicare surgery is d and 0 otherwise.

5.4.2 Simulation for scenario generation

5.4.2.1 Surgery steps and times:

Similar to Batun et al. [2011] we divide the surgery durations into 3 components: pre-incision, incision to closure, and post-closure activities. Pre-incision time involves

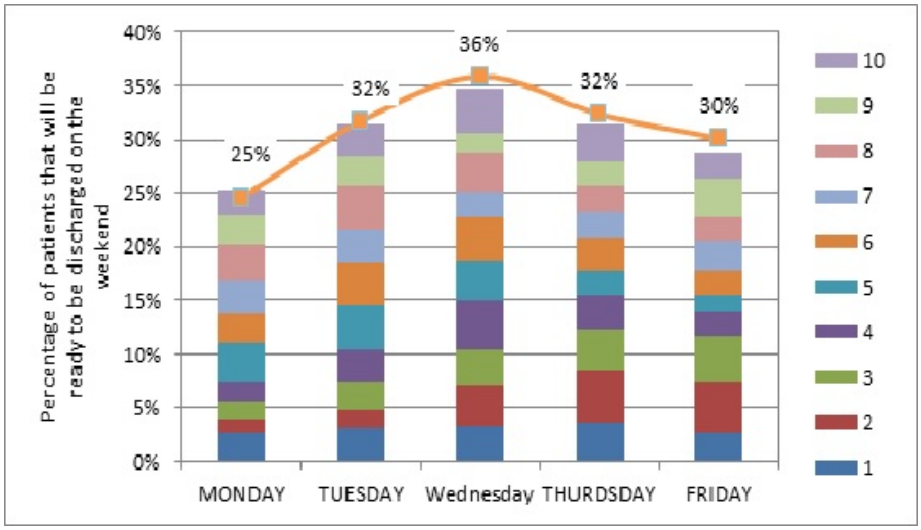


Figure 5.3: The percentage of weekend overflow by each patient category and surgery DOW

preparing the patient for surgery, incision to closure time is the actual procedure time and the post-closure is required to close up the incision and prepare the patient for recovery. Surgeons only need to be present in the OR for incision to closure; the other activities can be performed with other surgical staff present.

In addition to the pre-incision, incision to closure and post-closure time, we analyzed the surgeon turnover, OR cleaning and BOD time distributions as well (see Figure 5.4). Table 5.2 summarizes the best theoretical distribution fit of the empirical data. The lognormal function typically fits the best and is commonly used in the literature to represent similar highly variable procedure times (Spangler et al. [2004], Choi and Wilhelm [2012]). Additional information on these time-stamps is provided in Appendix F.

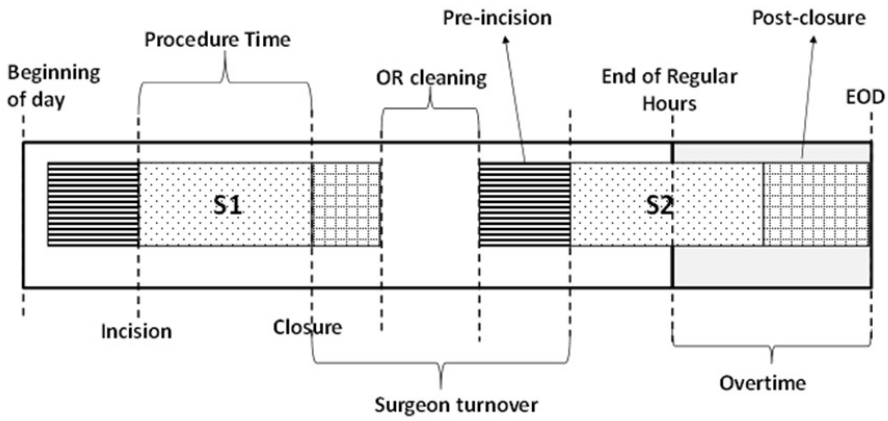


Figure 5.4: Stages in OR time

Table 5.2: Distributions for the time-stamps

	25% quartile	Median	75% quartile	Mean	Std dev	Best Fit
OR enter to incision [h]	1.1	1.4	1.7	1.4	0.47	Johnson Su
Incision to closure [h]	2.2	3.7	5.7	4.2	2.77	Weibull
Closure to OR exit [h]	0.3	0.45	0.6	0.5	0.29	Johnson Su
OR Turnover Time [m]	37	44	55	48.5	18.3	Normal 2

Due to the significant stochasticity of the problem environment we created a simulation model that mimics the surgical flow. The outputs of performing different surgical combinations were derived from the simulation and then used as inputs to the optimization model. The simulation used data for the time distributions of the 10 surgery categories. The distributions are derived for: beginning of day (BOD), pre-incision, incision to closure, post-closure, surgeon turnover, and OR cleaning for the 10 categories.

5.4.2.2 Why did we use a simulation model for outcomes projection?

Instead of using the empirical values for the surgical combinations, we have created a simulation model because not all the combinations were represented in the data-set. We were able to derive the results of interest (overtime, normalized NOI, utilization) using a simulation model for all possible surgical combinations, both for single surgeries, 2-surgery and 3-surgery pairs.

We have analyzed the convolution of lognormal variables, in order to predict the EOD for different surgical combinations. For instance, Gao et al. [2009] study the asymptotic behavior of a probability density function for the sum of any two lognormally distributed random variables. They approximate both the left and right with some simple functions. However, these models get intractable when we are considering the tail probability density function of more than two lognormally distributed random variables. Thus, we have turned our focus to using simulation models that mimics the ORs in Mayo Clinic based on sampling from historical data.

5.4.2.3 Parameters and scenarios:

We do not model the sequencing of surgeries, thus initially only 55 multiple surgery combinations are created. For example, from our modeling perspective 1 after 2 (1_2: a category 1 surgery followed by a category 2 surgery in the same OR) and 2 after 1 (2_1: a category 2 surgery followed by a category 1 surgery in the same OR) will result in the same EOD distribution. However, there were only a limited number of double and triple surgeries that could be performed, since most of the cases resulted in 100% overtime as can be seen from Figure 5.5. The red line represents the average

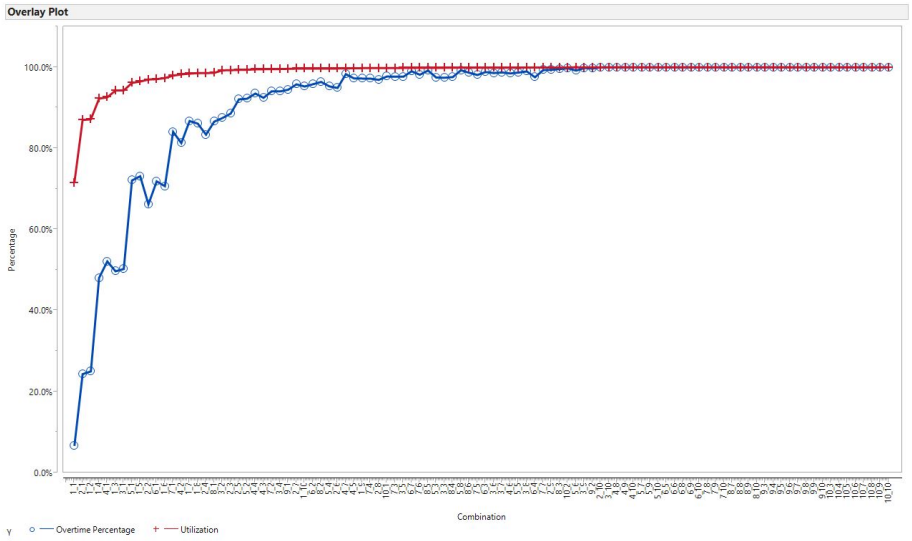


Figure 5.5: Percentage of overtime and utilization by patient category

utilization, whereas the blue line is the percentage of overtime after 5 PM. After our discussions with the surgeons, in all we came up with 42 surgical combinations (10 of which are individual surgeries performed in one day). However, for the case study we have only considered 20 combinations that result in feasible overtime.

We have compared the empirical EOD collected over 7 years with the results of the simulation model. Simulation model accurately predicts the EOD values of the empirical distribution, with 95% confidence, as can be seen from Figure 5.6. The cumulative distribution for the EOD values is always in between the two confidence intervals, almost always indistinguishable from the simulation values.

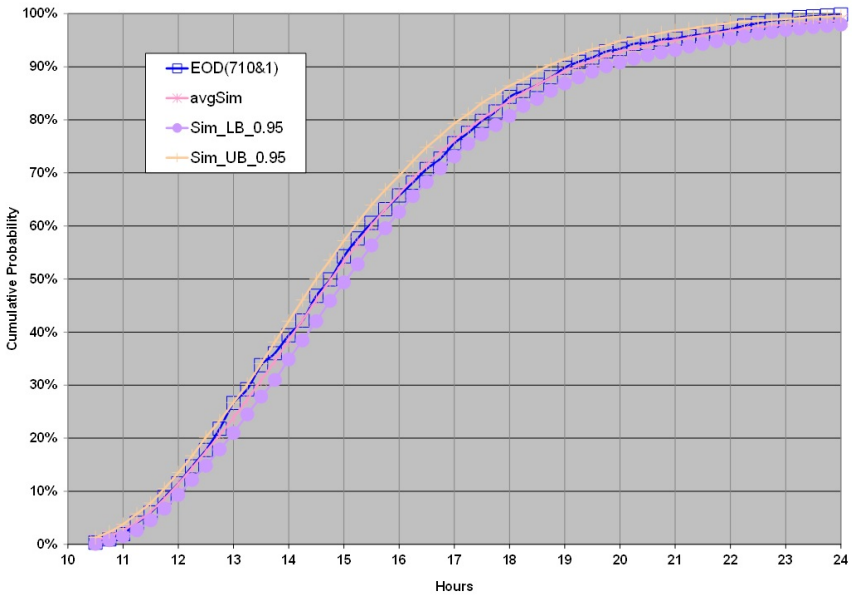


Figure 5.6: Validation of simulation model with 95th percentile, where the orange and purple lines represent the bounds for confidence interval.

5.4.2.4 Inputs to the optimization model:

Table 5.3 is the main input for the optimization model. It provides the output measures for all of the surgery combinations that are feasible for implementation. The values in parenthesis represent the Monday outputs, since Mondays start late.

5.4.3 Example optimization results

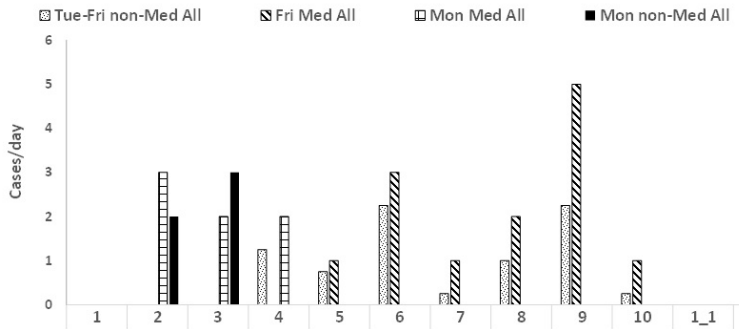
The outputs of the simulation model was used used to test and evaluate the optimization model. In addition, the optimization model was explored to consider tradeoffs and relationships among utilization levels, financial performance, overtime allowance, and case mix.

Table 5.3: Simulation inputs to the optimization model

Combination	EOD	% Days End After 5 PM	Hours After 5 PM	Utilization	Norm. NOI for Medicare	Norm. NOI for NON-Medicare	% overtime (11 PM)
1	10:18 AM	0.0 (0.0)	0.0 (0.0)	21.6 (31.6)	22%	41%	0.0 (0.0)
2	12:18 PM	0.9 (1.2)	0.0 (0.1)	40.7 (50.9)	25%	38%	0.1 (0.4)
3	1:24 PM	1.7 (4.4)	0.1 (0.1)	52.8 (62.8)	23%	42%	0.0 (0.2)
4	1:30 PM	5.2 (9.6)	0.1 (0.1)	53.9 (63.3)	20%	48%	0.0 (0.1)
5	2:54 PM	16.2 (29.8)	0.2 (0.4)	67.1 (76.5)	22%	45%	0.0 (0.1)
6	2:42 PM	12.1 (16.8)	0.3 (0.4)	64.0 (73.0)	20%	56%	2.0 (1.6)
7	4:06 PM	34.6 (44.3)	0.7 (1.0)	76.1 (82.3)	6%	87%	0.9 (1.8)
8	4:00 PM	23.8 (37.5)	0.7 (0.9)	74.6 (83.1)	19%	65%	3.9 (3.6)
9	6:06 PM	59.1 (74.4)	1.6 (2.2)	89.0 (94.0)	10%	77%	7.8 (8.8)
10	6:30 PM	66.6 (76.9)	2.0 (2.5)	90.9 (94.5)	13%	75%	2.1 (6.3)
1_1	2:18 PM	5.2 (14.0)	2.2 (1.5)	68.9 (78.4)	16%	54%	0.1 (0.6)
1_2	4:12 PM	18.6 (68.0)	2.0 (1.6)	85.5 (94.5)	19%	51%	1.7 (2.1)
1_4	5:24 PM	43.4 (80.5)	2.2 (2.9)	91.1 (96.6)	13%	61%	3.3 (5.3)
1_3	5:24 PM	42.9 (90.7)	1.8 (2.2)	93.5 (98.7)	17%	55%	2.6 (3.1)
2_2	6:06 PM	61.4 (96.8)	2.0 (2.8)	96.6 (99.6)	22%	47%	4.0 (4.9)
1_6	6:42 PM	66.3 (97.6)	2.7 (3.5)	96.9 (99.7)	14%	69%	7.5 (13.3)
1_5	6:48 PM	70.4 (94.0)	2.8 (3.4)	96.1 (98.8)	16%	58%	5.6 (14.6)
2_4	7:30 PM	81.7 (97.8)	3.2 (3.6)	98.4 (99.7)	16%	58%	10.8 (14.0)
1_8	7:54 PM	84.0 (94.9)	3.5 (4.1)	98.3 (99.3)	13%	78%	13.7 (16.7)
1_7	8:18 PM	86.7 (99.2)	4.1 (4.3)	98.3 (99.9)	0%	100%	21.3(18.2)

5.4.3.1 First stage optimization:

The optimal surgery mix determines the optimal values of proportion of overtime after 5 PM, expected hours of overtime after 5 PM, OR utilization levels, overtime percentage after 11 PM, normalized NOI and access (number of surgeries performed in comparison with the current access). As mentioned earlier, these reflect different stakeholders' perspectives. The optimal surgery mix is highly dependent on the



	1	2	3	4	5	6	7	8	9	10	1_1	2_1
Tue-Fri non-Med All	0	0	0	1	1	2	0	1	2	0	0	1
Fri Med All	0	0	0	0	1	3	1	2	5	1	0	1
Mon Med All	0	3	2	2	0	0	0	0	0	0	0	0
Mon non-Med All	0	2	3	0	0	0	0	0	0	0	0	0

Figure 5.7: Optimal surgery mix

constraints and the values of the parameters (overtime limit, proportion of case-mix bound, Medicare patient proportion, and planning horizon).

An example of the optimal surgery mix can be observed in Figure 5.7. The figure displays the optimal number of surgeries from each combination performed on different day groups. As can be seen, the shorter surgeries (which are the lower numbered categories) are performed on Mondays. Fridays are heavily loaded with longer Medicare patient procedures. This is intuitive, since the Mondays start late and in order not to create excessive overtime, long Medicare surgeries are left to Fridays, creating a greater surgery burden on these days. Note that for profitability reasons, Medicare patient surgeries are generally best scheduled on Mondays and Fridays (however not for all surgery types).

5.4.3.2 Second stage optimization:

The second stage creates the optimal 12 week schedule with the focus on maximizing the availability for multiple days staged surgeries. For example, the most common

occurrence of MDSS, a category 6 followed by a category 8, happens 16% of the time. The schedule assigns more priority to surgeries that have a greater empirical percentage. The schedule repeats itself every 12 weeks.

5.4.3.3 Third stage optimization:

Table 5.4 is an illustration of the final output of our optimization model, for one set of parameter values. This specific schedule is created so that it follows the blue-orange schedule template of Mayo Clinic. In the schedule, a blue week represents Surgeon 1 operating on Monday, Wednesday and Friday and Surgeon 2 on Tuesday and Thursday and an orange week represents Surgeon 2 operating on Monday, Wednesday and Friday and Surgeon 1 on Tuesday and Thursday. This stage ensures there is a balanced workload between blue and orange surgeons.

Table 5.4: Optimal 12 week blue-orange schedule, where blue week represents Surgeon 1 operating on Monday, Wednesday and Friday; Surgeon 2 on Tuesday and Thursday and orange week represents Surgeon 2 operating on Monday, Wednesday and Friday; Surgeon 1 on Tuesday and Thursday. The surgery combination each of the two surgeons will perform on each day for the 12 weeks is indicated.

Week Number	**Type of Week**	**Monday**	**Tuesday**	**Wednesday**	**Thursday**	**Friday**
W01	Blue Week	**2**	9	**9**	6	**7**
W02	Orange Week	3	**5**	8	**2_1**	6
W03	Blue Week	**3**	6	**4**	9	**6**
W04	Orange Week	2	**9**	2_1	**6**	2_1
W05	Blue Week	**2**	5	**4**	8	**6**
W06	Orange Week	4	**9**	4	**9**	10
W07	Blue Week	**2**	8	**8**	4	**6**
W08	Orange Week	3	**10**	7	**5**	9
W09	Blue Week	**4**	6	**9**	9	**8**
W10	Orange Week	2	**6**	9	**9**	6
W11	Blue Week	**3**	9	**9**	4	**5**
W12	Orange Week	3	**6**	9	**8**	6

5.4.3.4 Simulation for testing the robustness:

Even though, most of the spine patients tend to be pre-scheduled, there are also some urgent cases. Six percent of the time patients present infections, which need to be operated quickly (within 24 hours). These surgeries generally result in overtime, because infection patients need to be operated as the last surgery of the day, due to medical reasons (to prevent the spread of infections). The urgent cases typically take much shorter than regular surgeries (with an average length of 2 hours). Also, anecdotally on average 5% of the time last minute cancellations happen when the insurance company declines the surgery or when the health of the patient deteriorates.

We developed a second simulation model to test the impact of unplanned surgeries (infections) and cancellations. We analyzed the impact of these on EOD when utilizing the optimal schedule. We conclude that the simulation models and the results of our optimization model are robust and are not statistically different when compared with a year's worth of data (with a confidence interval of 99%).

5.4.3.5 Sensitivity analysis

We performed sensitivity analysis, in order to observe the impact of our constraints and associated parameters. We have changed the weight assigned to utilization in the objective function, the values of case-mix bound width, limit on overtime, length of planning horizon and analyzed the impact on optimal case-mix, NOI, expected overtime, total number of surgeries and utilization. We have used multi-variate analysis in order to study the interactions. This analysis has shown that planning

horizon did not have a statistically significant impact on any of the output measures. Thus we focused our sensitivity analysis on the impact of bound-width, overtime limit and weights.

In order to understand the trade-off between NOI and utilization, we generated an efficient frontier (as can be seen in Figure 5.8). We altered the weights assigned to these two output measures, to generate the non-inferior points curve. We have performed this analysis for different bound widths and overtime limits. It is possible to gain more while utilizing the ORs the same level, but by changing the patient mix. The initial flat line in the curve shows the potential gain in NOI without sacrificing from utilization. The underlying reason is that, the surgeries that are creating high utilization levels do not necessarily result in higher revenue.

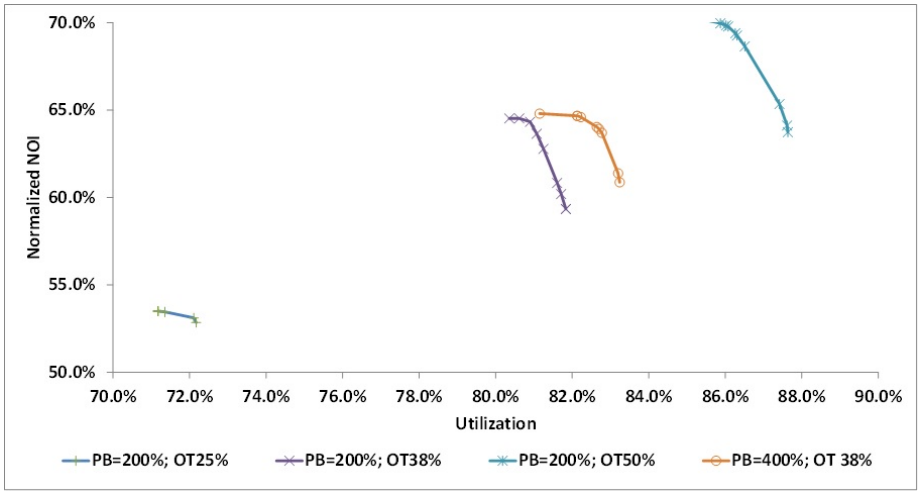


Figure 5.8: Trade-off between utilization and NOI

We have used this sensitivity analysis in order to set the values of the parameters in the optimization model for the pilot study implementation. We have presented the results of our sensitivity analysis to our stakeholders and discussions about the

tradeoffs have led to the values of parameters used for the pilot study. Some of the main parameter values are: w (weight assigned to utilization)= 80%, o (overtime percentage after 5 PM)= 25%, e (number of hours past 5 PM)= 5 hours, f (percentage of days that end after 11 PM)= 5%, m (Medicare patient proportion)= 30%, T (planning horizon)= 120 days. We provide additional research on sensitivity analysis in Appendix G.

5.5 Implementation

The optimized scheduling approach was implemented via a custom designed web-based application that partially integrates with Mayo Clinic's existing surgical planning systems. The application, Spine Surgery Scheduling Optimization (SSSO), provides visual cues to promote scheduling surgeries on the appropriate days identified by the optimization model. If a surgeon or their delegated scheduler needs to schedule a case on a "non-optimal" day, the tool provides visual information as to the case load and the likelihood of going overtime. The application can be used on any office or tablet computer and is therefore easy to use in an interactive way with the patient. Figures 5.9, 5.10 and 5.11 show screenshots of the web-based application.

To evaluate the effectiveness of the optimization model and SSSO application, a pilot study was run from December 2012 to June 2013. Two of the four orthopedic spine surgeons participated in the study. It should be noted that other initiatives were going on at the same time as the pilot. In particular the orthopedic spine practice was working to increase case volumes and improve work processes related to on-time case starts and room turnover. Therefore, as in an intervention to an on-going process,

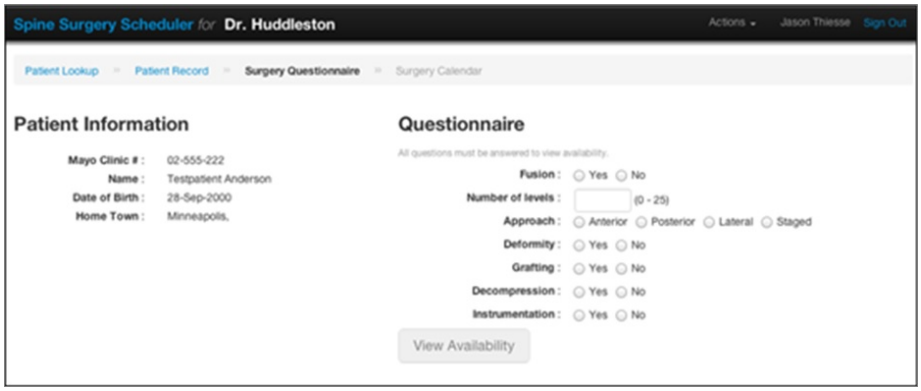


Figure 5.9: SSSO screenshots

it is difficult to determine the precise benefit or cost of the implementation. In the following section we will describe the results of the pilot and how we attempted to account for process effects not due to SSSO.

5.5.1 Results of the pilot implementation

In evaluating the results, the first month of the pilot data was removed, because surgical cases during this period were primarily scheduled using the old approach. Figure 5.12 shows the results for the key performance measures during the evaluated pilot period. For all measures, we consider only days during which surgeons had cases scheduled, thus we eliminated empty days that were due to holidays, vacations, and on-call duties. For utilization, this was evaluated as the busy percentage of the prime time period of 7:30 AM to 5:00 PM. Overtime is defined as the percentage of days that went over 5 PM.

Figure 5.12 shows that, in general, the implementation of the SSSO system provided the desired results. Patient access and utilization were higher and overtime

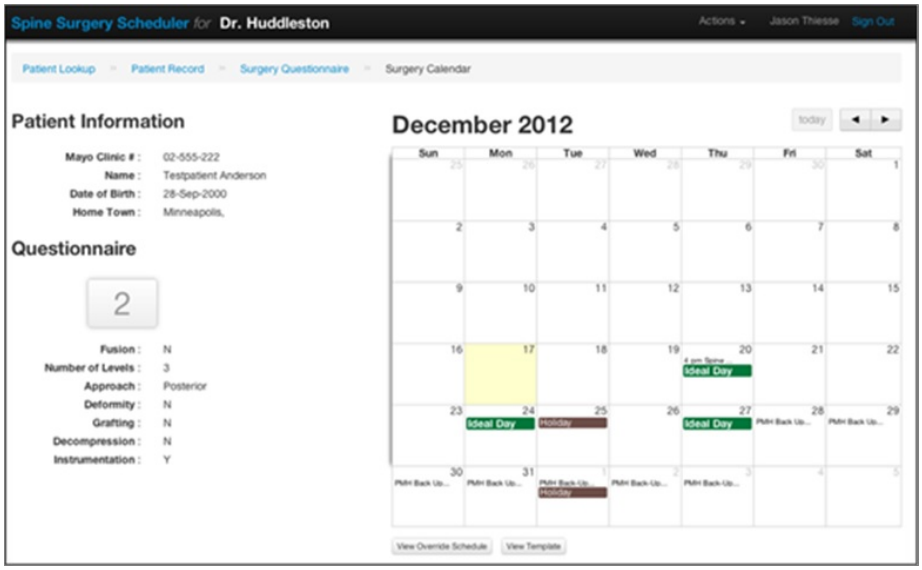


Figure 5.10: Question screen to categorize surgical case

Figure 5.11: Initial screen that identifies optimal days

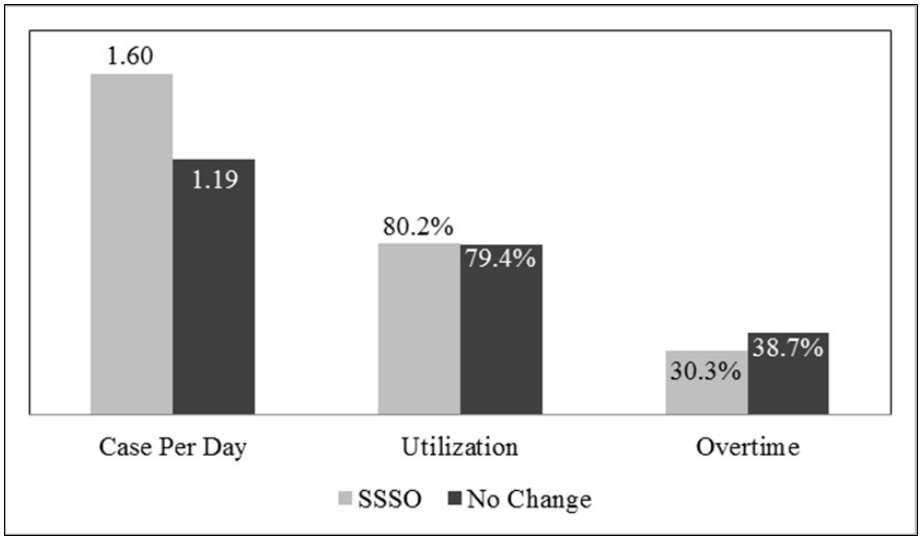


Figure 5.12: Comparison of output measures evaluated during the pilot

lower for the surgeons participating in the pilot. In particular, it is interesting to note the significant increase in cases per day for the surgeons participating in the pilot. It is also important to note that this was not done by using more overtime.

Of course, it may be that the surgeons participating in the pilot had practices that performed better before the pilot. Therefore, we also present the pre and post-implementation results for all surgeons in Table 5.5.

It is gratifying to identify that the overall efforts of the practice to improve patient access were achieved because all surgeons increased their number of cases per day during the pilot evaluation period. The two surgeons participating in the SSSO pilot increased their access by a higher percentage (30.1%) versus the non-participating surgeons (24.6%). In our study, Surgeon 1 achieved the kind of results the optimization method was intended to return: an increase in cases per day, prime-time utilization, and decrease in days going to overtime. Surgeon 2, who was also

Table 5.5: Pre and post-implementation results for all surgeons

	Cases per day		Utilization		Overtime	
Surgeon	Pre-Imp	Post-Imp	Pre-Imp	Post-Imp	Pre-Imp	Post-Imp
S1	1.30	1.57	72%	77%	30%	24%
S2	1.16	1.63	70%	83%	21%	37%
S3	1.07	1.30	75%	83%	33%	48%
S4	0.85	1.09	61%	76%	25%	29%

Surgeons 1 and 2 participated in the SSSO pilot implementation.
Shaded values show statistical difference at 0.05 significance level.

involved in the pilot increased access and utilization, but also significantly increased days going to overtime. Thus, we would say that Surgeon 1 used a "working smarter" approach and Surgeon 2 a "working harder" approach (as it appears Surgeon's 3 and 4 also did). Working hard is, of course, commendable, but the continued strain on the surgeons and the surgical teams working in this mode may not be sustainable or safe in the long run.

From a financial perspective, the average per case increase in NOI for the two surgeons participating in the pilot was 122%. For the non-participating surgeons the average actually decreased by 25%. As part of this analysis we also considered if there were any significant changes in the mix of patients by payer type. For the surgeons participating in the SSSO pilot there was a small overall drop in the proportion of government paid patients, however, the proportion was still well above the minimum established by the practice. Also, profitability for Medicare patients increased during the pilot period, suggesting that the efforts to do these surgeries on the best days to avoid uncompensated hospital days were effective. Together with the overall increase in access due to the SSSO implementation and other improvement initiatives, the

financial sustainability in the orthopedic spine surgery practice has improved and will provide better access for all patients, regardless of reimbursement type, in the future.

In summary, the pilot implementation was deemed successful, but not as comprehensively as desired. As the pilot rolled out, several challenges occurred, including technical issues with the programming of SSSO, lack of desired flexibility in scheduling patients, and some discomfort by users of the tool with its reliability. The following section will discuss some of the lessons we learned and proposed solutions as the system is rolled out more broadly across the surgical practice at Mayo Clinic.

5.5.2 Lessons learned from the pilot

Pilot implementations by their nature are intended as learning experiences. The points below are some of the key lessons we learned from our pilot.

- The SSSO application was generally developed in the classic waterfall approach. The optimization team handed off a completed method to the programming team. There was some integration and communication, but not as much as desired. This resulted in some technical issues with the tool. Some of these issues were the responsibility of the optimization team and some the responsibility of the programming team. All or most of these issues could have been avoided by earlier involvement and better integration of the teams.

- Some assumptions were built into the optimization method that did not work in practice. In particular, we assumed that case-mix could be shaped by how access was controlled at the time of surgery scheduling. However, for spine

surgery it is common for the surgeons to see patients several times before the surgery decision is made. Limiting a patient's surgical access when they had developed a relationship with a surgeon pushed against Mayo Clinic's high-quality service philosophy. Thus, this approach is being adapted for ongoing implementations. Efforts at controlling access before patients come to Mayo Clinic have been implemented and are still under way that will ensure the best use of our capacity while ensuring the needs of the patient come first.

- As identified in the previous section, surgeon 1 had the most desired performance profile during the pilot. This surgeon and his scheduling team were the most involved during the optimization and tool development process. It is not surprising that the staff in this group had the most confidence in and understanding of what the tool was trying to accomplish. For ongoing implementations of the modified tool we are working to involve more surgeons and staff in the development process.

- Both surgeons in the pilot found the ability to see the impact of scheduling a particular case on a day, even if they were overriding what was recommended by the optimization. The visualization shown in the window in Figure 5.11, was of particular value. As scheduling decisions evolve from being very patient preference oriented to being more system optimized, providing the surgery schedulers with useful information to guide decision making with flexibility is being incorporated into new versions of the tool.

A great deal was learned from the pilot and specific improvements are being incorporated into new versions that are in development for several surgical practices at Mayo Clinic.

5.6 Conclusions

In this paper we presented an improved method for scheduling spine surgeries in the orthopedic spine surgery practice at the Mayo Clinic. The method we developed addresses specific elements of spine surgery at Mayo Clinic, however, the general concepts used to develop the method are likely to be useful at other healthcare organizations. Unique aspects of our model include the incorporation of both resource utilization and financial objectives. The latter was also addressed by considering the profitability of the patients entire encounter related to their surgery, including post-surgery hospitalization and the effects of unnecessary hospital stays (and associated costs) for patients likely to require skilled nursing facilities upon discharge. Further, categorizing surgeries and developing statistical models for predicting surgical lengths using clinical factors is a key contribution. Using input from the surgeons to categorize case types that led directly to scheduling decisions assisted in gaining clinical staff engagement.

An implementation using a customized web-based tool that incorporated our optimization model showed generally positive results. Patient access improved significantly for the surgeons involved in the pilot and operating room utilization improved marginally. For one of the two surgeons participating in the pilot the access benefits were achieved by also reducing the percentage of overtime days. It should be noted

that patient access also increased for the surgeons not participating in the pilot, but not by as much.

There are some topics we have failed to address in this paper. We consider hospital LOS implicitly in considering profitability, but the impact on downstream resources is not investigated. We consider the surgeons as bottlenecks and the impact on inpatient or PACU beds is not in the scope of this project. In general, at Mayo Clinic in Rochester, these resources are not constraints. Lastly, due to lack of information about cancellations, we did not directly incorporate these into our model.

Since the surgical durations are relatively long, the number of surgical combinations is restricted (20 surgical combinations). As the number of surgery pairs increase (for specialties with shorter durations) the computational burden will increase as well. We have developed the optimization model using both Excel Solver and AMPL. Excel was favored for implementation and the pilot study, and the computational time was around an hour for each stage. AMPL, which should be favored for research, on the other hand, provides solutions in less than 5 minutes for each stage. Exploring the general problem (with a greater number of decision variables) will allow us to understand the computational complexity of the optimization model more accurately.

Thus, while this chapter highlights a specific case study application, we believe that many of the results and insights will be of interest more broadly. In particular, the emphasis on considering the tradeoffs and effects of constraint limits may help other similar surgical operations gain useful insights. At Mayo Clinic the gen-

eral approach we developed is being considered for other surgical services and would likely benefit other organizations. Thus, while our paper discusses the specific implementation, we emphasized the underlying ideas and theory of the application and show results of experiments that develop managerial insight. Other surgical services such as cardio-thoracic, neurosurgery, and plastic surgery that have long average and highly variable procedure times may benefit from our research as well. As reported in Abouleish et al. [2003] these services together (with spine surgery) may make up to about 20% of surgical volume in hospitals.

From a literature perspective we believe our paper is a significant contribution because it does more than just consider the issues of changing case-mix and surgical scheduling (which are prevalent in the conceptual operations management literature). We extend the research area by considering the multiple objectives related to utilization (and correspondingly, patient access), overtime, and financial performance. Further, considering the downstream financial issues related to an important class of patients (those with fixed reimbursements) is novel and increasingly important, particularly in the U.S. where healthcare reform is a prominent issue. Finally, considering the behavioral aspects of the patients and those doing the surgical scheduling is unique.

CHAPTER 6

FUTURE WORK

In this chapter, I outline the research plan that I will conduct in the future. The material in this chapter is organized under the three application areas.

6.1 Opportunities in Primary Care

6.1.1 Testing the applicability of the findings for primary care on a national level

The National Ambulatory Medical Care Survey (NAMCS) is a national survey that provides data on ambulatory medical care services in the United States. Also Medical Expenditure Panel Survey (MEPS) collects nationwide data on the health services that Americans use, how frequently they use them, the cost of the services and the insurance that the patients have. By using these datasets, we can test if the heuristics we developed to improve timeliness and continuity in group practices using data from Mayo Clinic, would create the same significant impact at the national level.

6.1.2 The Patient Centered Medical Home (PCMH) in primary care

Patient-centered medical homes (PCMH), a new model of primary care delivery, is a research line to explore. PCMH aims to reorganize primary care to improve access, coordination, quality, satisfaction, and provide comprehensive patient-centered care (Nutting et al. [2009]). This has a nationwide importance since 15 to 24 million additional primary care visits are expected as a result of the increase in demand from Affordable Care Act (ACA) (Hofer et al. [2011]). Compounding the increase in patient volumes and the shortage of primary care workforce, is the aging population and the epidemic of chronic diseases, which will likely give rise to more patients with multiple comorbidities, requiring more PCP time and resources. Currently, 45% of the U.S. population has chronic conditions requiring care management. Of this population, 60 million, or roughly half of those with chronic conditions, have multiple conditions (Kopach-Konrad et al. [2007]).

Capacity design of a PCMH is challenging, since care coordination across multiple providers, email, phone and home visits need to be considered. Impact of non-visit care is an important topic in medical homes. Patients require care outside of office visits, much of which is not reimbursed. Non-visit care activities include emails, telephone calls, refilling prescriptions, reviewing consultations, lab test results and imaging reports. Studies show that almost one half of PCPs' workday involves these non-face-to-face tasks (Chen et al. [2011]). On the one hand, these may improve access and reduce office visits, however, these are a huge burden on PCPs' workload and are not reimbursed (Dyrbye et al. [2012]).

Modeling these non-face-to-face tasks in calculating the optimal panel size and case-mix is an essential extension to our paper. Otherwise the calculations might be misleading and potentially underestimate the workload of a physician. This stochastic capacity allocation problem is further complicated by the patient preferences.

The National Ambulatory Medical Care Survey (NAMCS) and Medical Expenditure Panel Survey (MEPS) can again be used in order to start investigating the impact of patient characteristics (patient mix) and preferences on the design of medical homes.

6.1.3 Studying the relationship between readmissions and access to primary care

Studying the relationship between readmissions, access to primary care and the number of comorbidities is an interesting research direction. Hospital readmission rates have become an important predictor for both quality and costs. This is partly because of the very high readmission rates (17.6% of Medicare patients were readmitted within 30 days, resulting in $15 billion annually), but more importantly 10-50% of readmissions are potentially avoidable (MedPac [2007]).

Discharges can be looked at as a transfer of the responsibility of care from the hospitalist to the primary care provider. Traditionally, PCPs admitted their own patients, provided hospital care and followed them after their discharge. However, since this became unsustainable over the years, the hospitalists have started to take care of the hospital medicine side and PCPs the outpatient side (Wachter and Goldman [1996]). However, this discontinuity in care hinders the PCPs' ability to provide adequate follow-up care, increasing the risk of a readmission (Harding [2002], Kripalani

et al. [2007a]). This is especially because of the shortfall in the communication of information between the hospitalist and the PCP. The most typical way of communication is through the discharge summaries, which generally fail to provide important information on patients' medical condition. Also surveys show that typically these summaries do not arrive to PCPs on time for the follow-up appointment of the patient (Kripalani et al. [2007b], van Walraven and Weinberg [1995]).

Relatively few studies have looked at the relationship between primary care and readmissions. Jencks et al. [2009] show that one in five Medicare patients ends up back in the hospital within 30 days, and of those readmitted within 30 days, 50% did not see their PCP for a follow-up appointment after their hospital discharge. It is essential to analyze if this finding can be extended to non-Medicare beneficiaries as well. Weinberger et al. [1996] find conflicting results for veterans discharged from Veterans Affairs hospitals, in their study to test the impact of a primary care intervention. This intervention was designed to improve the veterans access to primary care providers, which actually increased the rate of re-hospitalization. However, patients in the intervention group were more satisfied with their care. On the other hand, Bodenheimer and Pham [2010] show that a good chronic care management can significantly decrease hospital readmissions for certain types of chronic conditions. In support of this point, Donz et al. [2013] study higher risk groups for readmissions and show that patients with certain chronic conditions like heart failure, and chronic kidney disease have a higher risk of readmission.

It is a thought-provoking research field to study whether certain comorbidities have a higher probability of readmission or not, and what kind of an action plan

can be developed accordingly. The post-discharge care of the patient will need to consider the risk factors that might lead to a higher probability of readmission, and not only the acute condition of the patient. More attention could be given to these patients for a closer follow-up to monitor their chronic conditions. More importantly, a better coordination of care between the hospitalists and the PCP will be especially crucial for these types of "high-risk" patients. This idea again relates back to the PCMH structure, which addresses each patient's unique needs and also values a clear communication and coordination across patients, the medical home, and members of the patients' healthcare team (Rittenhouse et al. [2009]). Hospitalists and a tighter connection between PCPs and hospitalists, are crucial for a successful implementation of a PCMH (Collins [2012b]).

6.2 Opportunities in Inpatient Bed Planning

Our inpatient bed planning project can be potentially extended in the following areas:

6.2.1 Pre and post allocation delays

As a possible extension, we aim to integrate pre and post bed allocation delays to model the possible secondary bottlenecks, including but not necessarily limited to staff shortages. In reality even if a bed is available for the patient, a patient can experience a pre-allocation delay first, and then a post-allocation delay before being transferred to an inpatient bed (Shi, P. [2013]). This would mean explicitly modeling the operational delays that are caused by resource constraints (like ED and

unit nurses) other than bed unavailability in inpatient units. Thus the time that the bed is available is not necessarily the same as the time that patient is in the bed. This could be integrated into the model by adding a delay (either deterministic or based on a probability distribution) for each patient based on the hour of the day.

6.2.2 Modeling readmissions based on different discharge policies

As discussed in Section 6.1.3, readmission refers to a patient being admitted to a hospital within a certain time period from the initial admission. In the Medicare framework, readmissions happen when a patient is being hospitalized within 30 days of an initial hospital stay. There are many factors that affect the possibility for readmission, including patients' diagnoses and severity; patients' behavior and the quality of post-discharge care (James [2013], Kripalani et al. [2007b]). Thus, some patients are more prone to readmissions than others and the studies show that the discharge planning has a major impact on the probability of a readmission. The readmissions perspective can be integrated to our simulation framework by including a readmissions probability based on different discharge policies or the type of patients.

6.2.3 Incorporating uncertainty to the discharge process

In our simulation model, the discharge process is assumed as deterministic, so once the LOS of the patient is complete (which was randomly sampled from empirical distribution) the model assumes the patient is ready to leave the hospital. However, in reality (which is in fact reflected in the data implicitly), even when the patient is considered as a potentially dischargeable patient for the next day, there is a prob-

ability that the patient will not be able to go home. This can be due to a couple of reasons: a healthcare delay (deterioration in patient's health), a problem related to the availability in post-acute care facilities (like skilled nursing facilities), inconvenient discharge timing for the family members, or a social care delay. Thus there can be many factors that influence the timing of the discharge process and make it random.

In order to make the discharge process stochastic, we can assign a probability distribution based on a patient's type. These probability distributions can be derived from the expert opinion of our nursing collaborators.

6.3 Opportunities in Spine Surgery Scheduling

6.3.1 Modeling other types of uncertainties

The models presented in this study consider only the uncertainty in surgery durations. Depending on the characteristics of the surgical specialty and the healthcare institution, other types of uncertainties such as add-on surgeries, patient no-shows or cancellations can also be included in the optimization model. We have tested the impact of these factors on our optimization model using a simulation. However, a more realistic approach which incorporates both no-shows and the dynamic nature of the appointment scheduling process can be developed based on our current models.

6.3.2 Extended surgery scheduling model in the presence of other resources and uncertainty

We have only considered the surgeons and operating rooms as bottlenecks in our surgery scheduling model. An updated model could include multiple stages of the

hospital service (e.g. surgery followed by recovery). In other words, linking the unit census levels to an operating room schedule and/or other critical hospital subsystems (surgical ICUs and PACUs) can be a potential extension. With more general models, the simultaneous effects of demand uncertainty from no-shows and add-on cases could be better estimated.

6.3.3 Testing the robustness of the model by extensions to other surgical services

At Mayo Clinic the general approach we developed is being considered for other surgical services (e.g., Neurosurgery) and would likely benefit other organizations. Thus, while our paper discusses the specific implementation, we emphasize that the underlying ideas and theory of the application can be used to develop managerial insight.

Other surgical services such as cardio-thoracic, neurosurgery, and plastic surgery that have long average and highly variable procedure times may benefit from our research as well. As reported in Abouleish et al. [2003] these services together (with spine surgery) may make up to about 20% of surgical volume in hospitals. However, every specialty has its own set of constraints and objectives. Thus, this kind of implementation will require further data mining analysis specific to that specialty. This will include a new categorization model for grouping the patients with similar surgical durations and creating the required inputs to the optimization model. Testing the robustness of the optimization model with other surgical services can be a potential next step.

APPENDIX A

QUEUING MODEL FORMULAS FOR CHAPTER 3

Formula for the M/M/1 Queue

W_q(average waiting time in the queue)= $\frac{\lambda}{\mu(\mu-\lambda)}$

Formula for the M/M/2 Queue

ρ(utilization)= $\frac{\lambda}{2\mu}$

$W_q = \frac{\lambda^2}{\mu(1-\rho^2)}$

Formulas for the Partial Pooling Model

We have directly used the approach described in Guo and Hassin [2012], which is summarized in this section, in order to derive numerical results.

For the symmetric case:

As described in Section 3.2.1, λ_1 is the arrival rate of the dedicated patients of Physician 1, λ_3 is the arrival rate of the dedicated patients of Physician 2. And λ_2 represents arrival rate of the flexible patients, which is equal to $\sum_{i=1}^{k} \lambda_i^0 * N_i$, where N_i is the total number of patients from category i and patients from category 1 to k are shared (λ_i^0 values represent the Mayo comorbidity count visit rates). x_{ij} values are used to calculate the values for λ_1 and λ_3 and it represents the number of dedicated patients from category i assigned to physician j. $\lambda_1 = \sum_{i=k+1}^{M} \lambda_i^0 * x_{i1}$ and $\lambda_3 = \sum_{i=k+1}^{M} \lambda_i^0 * x_{i2}$, where M is the total number of categories. And λ is the total arrival rate to the practice, thus $\lambda = \lambda_1 + \lambda_2 + \lambda_3$.

Guo and Hassin [2012] assume that Type 2 (flexible) customers who see both servers idle, will choose to join server 1 in probability κ and server 2 with probability $1 - \kappa$. And for all of the balance equations to hold, κ needs to satisfy:

$\kappa = \frac{\lambda_2+\lambda_3}{\lambda_1+2\lambda_2+\lambda_3}$

Thus, κ represents the chance that a customer will choose which idle server to join is proportional to the inverse the total patient arrival rate into that server's queue.

For the symmetric case, Guo and Hassin [2012] assume $\mu_1 = \mu_2 = \mu$, $\kappa = 0.5$ (flexible patients randomly pick up the two queues in equal probability) and $\lambda_1 = \lambda_3 = (\lambda - \lambda_2)/2$. Using these they derive the waiting times for the three patient types.

$W_1 = W_3 = \frac{1}{2}\frac{\lambda}{\mu}\frac{8\mu^2+4\lambda_2\mu+\lambda_2^2-4\mu\lambda-\lambda\lambda_2}{(2\mu-\lambda)(2\mu-\lambda+\lambda_2)(2\mu+\lambda_2)}$

$W_2 = \frac{1}{2}\frac{\lambda}{\mu}\frac{\lambda+\lambda_2}{(2\mu+\lambda_2)(2\mu-\lambda)}$

For the asymmetric case:

We create a variable p, which represents the percentage of flexible patients (a measure of continuity), calculated as $p = \frac{\lambda_2}{\lambda_1+\lambda_2+\lambda_3}$.

In order to perform the calculations for the asymmetric case, we also need to derive the values for λ_1', and λ_3'. λ_1' is the total number of patients Physician 1 cares for, similarly λ_3' is the total number of patients Physician 2 cares for, including the shared patients. First, we assume that the shared patients are assigned to physicians based on the κ proportions. Thus κ percent of the shared patients will see Physician 1, and $(1-\kappa)$ percent of the patients will see Physician 2.

$\lambda_1' = \lambda_1 + (\lambda_2 * \kappa)$

$\lambda_3' = \lambda_3 + \lambda_2 * (1-\kappa)$

The weighted service time for each surgeon is adjusted based on the values of λ_1' and λ_3' (as a result of change in the number of patients shared).

We have also altered the algorithm of how the shared patients were assigned to physicians and analyzed the impact of equally assigning the shared patients to physicians. For example if only 0 comorbidity patients are shared, we assume half of the 0 comorbidity patients see Physician 1 and other half sees Physician 2 (like in symmetric case). The values for λ_1', and λ_3' are calculated as follows:

$\lambda_1' = \lambda_1 + \frac{\lambda_2}{2}$

$\lambda_3' = \lambda_3 + \frac{\lambda_2}{2}$

The results were almost identical when using κ and assigning half of the shared patients. Thus, we only present one set of results for assigning half of the shared patients.

Methodology for Evaluating the Waiting Times:

We now present the steps we followed for evaluating the waiting times for λ_1, λ_2 and λ_3 customers.

1. Define the state variable age time, which is the total time spent by the oldest customer in the queue.

$V_i(t)$: Age of customer waiting at the head of Q_i at t, if $V_i(t) \geq 0$; -Time from t until next arrival instant, if $V_i(t) < 0$.

The sample path of the three-dimensional age process is defined with $V = (V_1, V_2, V_3)$

2. Define the positive age times as follows:

$X = V_1^+ = \max\{0, V_1\}$

$Y = V_2^+ = \max\{0, V_2\}$

$Z = V_3^+ = \max\{0, V_3\}$

And the system has three queues: Q1, Q2 and Q3.

3. Use the formula 25 in Guo and Hassin [2012] to calculate the value of F_0 (probability that all queues are empty):

$$F_0 = \frac{C}{(\lambda_1\lambda_2\lambda_3)}\frac{1}{(\lambda_1+\lambda_2+\lambda_3)}\left(\frac{(\mu_1\mu_2)}{(\lambda_1+\lambda_2)} + \frac{(\mu_1\mu_2)}{(\lambda_3+\lambda_2)}\right)\frac{C}{(\lambda_1\lambda_2\lambda_3)}\left(\frac{\mu_1}{(\lambda_1+\lambda_2)} + \frac{\mu_2}{(\lambda_3+\lambda_2)}\right) + \frac{C}{(\lambda_1\lambda_2\lambda_3)}$$

4. And use F_0 in Formula 26 to solve for the constant C from the normalization condition:

$$F_0 + \frac{(\lambda_2+\lambda_3+\mu_2)}{(\lambda_2+\lambda_3)}\frac{C}{(\lambda_2\lambda_3(\mu_1-\lambda_1))} + \frac{C}{(\lambda_1\lambda_3(\mu_1+\mu_2-\lambda_2))} + \frac{(\lambda_1+\lambda_2+\mu_1)}{(\lambda_1+\lambda_2)}\frac{C}{(\lambda_1\lambda_2(\mu_2-\lambda_3))} + \frac{C}{(\lambda_2(\mu_1-\lambda_1)(\mu_2-\lambda_3))} + \frac{C}{(\lambda_1(\mu_1-\lambda_2))}\left(\frac{1}{(\mu_2-\lambda_3)} - \frac{1}{(\mu_1+\mu_2-\lambda_2-\lambda_3)}\right) + \frac{C}{(\lambda_1(\mu_1+\mu_2-\lambda_2)(\mu_1+\mu_2-\lambda_2-\lambda_3))} + \frac{C}{(\lambda_3(\mu_1+\mu_2-\lambda_2)(\mu_1+\mu_2-\lambda_1-\lambda_2))} + \frac{C}{(\lambda_3(\mu_2-\lambda_2))}\left(\frac{1}{(\mu_1-\lambda_1)} - \frac{1}{(\mu_1+\mu_2-\lambda_1-\lambda_2)}\right) + \frac{C\mu_1\mu_2}{\lambda_1\lambda_3(\mu_1-\lambda_1)(\mu_2-\lambda_3)(\mu_1+\mu_2-\lambda_1-\lambda_2-\lambda_3)} - \frac{C\mu_1}{\lambda_1\lambda_3(\mu_1-\lambda_1)(\mu_1+\mu_2-\lambda_1-\lambda_2)} - \frac{C\mu_2}{\lambda_1\lambda_3(\mu_2-\lambda_3)(\mu_1+\mu_2-\lambda_2-\lambda_3)} + \frac{C}{\lambda_1\lambda_3(\mu_1+\mu_2-\lambda_2)} = 1$$

5. After obtaining the value of C use formulas on page 39 from Guo and Hassin [2012] to find the performance measures ($E(X)$ and $\Pr(X > 0)$).

a) Derive the mean aging times:

$E(X)$: The expected waiting time for an arrival to move to the front position of Q1.

$E(X) = \frac{\mu_2^2}{\lambda_2\lambda_3(\lambda_2+\lambda_3)(\mu_2-\lambda_2-\lambda_3)} * \frac{C}{(\mu_1-\lambda_1)^2} - \frac{\mu_1\mu_2}{\lambda_3(\mu_2-\lambda_3)(\mu_2-\lambda_3)(\mu_2-\lambda_2-\lambda_3)(\mu_1+\mu_2-\lambda_2-\lambda_3)} * \frac{C}{(\mu_1+\mu_2-\lambda_1-\lambda_2-\lambda_3)^2}$

b) Calculate the probability of the existence of a positive X value:

$Pr(X > 0) = 1-F_0-\frac{C}{\lambda_1\lambda_3(\mu_1+\mu_2-\lambda_2)}-\frac{C(\lambda_1+\lambda_2+\mu_1)}{\lambda_1\lambda_2(\lambda_1+\lambda_2)(\mu_2-\lambda_3)}-\frac{C(\mu_1+2\mu_2-\lambda_2-\lambda_3)}{\lambda_1(\mu_2-\lambda_3)(\mu_1+\mu_2-\lambda_2)(\mu_1+\mu_2-\lambda_2-\lambda_3)}$

c) Perform the same operation for Y.

$E[Y] = \frac{(\mu_1\mu_2)}{\lambda_1\lambda_3(\mu_1-\lambda_1)(\mu_2-\lambda_3)} * \frac{C}{(\mu_1+\mu_2-\lambda_1-\lambda_2-\lambda_3)^2}$

$Pr(Y > 0) = 1 - F_0 - \frac{C(\lambda_2+\lambda_3+\mu_2)}{\lambda_2\lambda_3(\lambda_2+\lambda_3)(\mu_1-\lambda_1)} - \frac{C(\lambda_1+\lambda_2+\mu_1)}{\lambda_1\lambda_2(\lambda_1+\lambda_2)(\mu_2-\lambda_3)} - \frac{C}{\lambda_2(\mu_1-\lambda_1)(\mu_2-\lambda_3)}$

6. Use Equation 27 to calculate W_1 (expected waiting time in the queue for a Q1 customer):

$W_1 = E[X] + \frac{Pr(X>0)}{\lambda_1}$

7. Use Equation 28 to calculate W_2.

$W_2 = E[Y] + \frac{Pr(Y>0)}{\lambda_2}$

Likewise, W_3 is calculated.

8. Q1 and Q3 customers have their own dedicated servers and their expected waiting times can be calculated as:

$S_1 = W_1 + \frac{1}{\mu_1}$

$S_3 = W_3 + \frac{1}{\mu_2}$

9. Expression E(S2) is derived as follows:

Guo and Hassin [2012] define P1 to be the probability of only 1 server being busy and P2 to be the probability of both servers being busy.

$P_1 = \frac{C}{\lambda_1\lambda_2\lambda_3}(\frac{\mu_1}{\lambda_1+\lambda_2} + \frac{\mu_2}{\lambda_3+\lambda_2}) + \frac{C\mu_2}{\lambda_2\lambda_3(\lambda_3+\lambda_2)(\mu_1-\lambda_1)} + \frac{C\mu_1}{\lambda_1\lambda_2(\lambda_1+\lambda_2)(\mu_2-\lambda_3)}$

$P_2 = 1 - \frac{C}{\lambda_1\lambda_2\lambda_3}\frac{1}{\lambda_1+\lambda_2+\lambda_3}(\frac{\mu_1\mu_2}{\lambda_1+\lambda_2} + \frac{\mu_1\mu_2}{\lambda_3+\lambda_2}) - \frac{C}{\lambda_1\lambda_2\lambda_3}(\frac{\mu_1}{\lambda_1+\lambda_2} + \frac{\mu_2}{\lambda_3+\lambda_2}) - \frac{C\mu_2)}{\lambda_2\lambda_3(\lambda_3+\lambda_2)(\mu_1-\lambda_1)} - \frac{C\mu_1}{\lambda_1\lambda_2(\lambda_1+\lambda_2)(\mu_2-\lambda_3)}$

Thus the expected waiting time for Q2 customers is:

$S_2 = W_2 + \frac{P_1+2P_2-\frac{\lambda_1}{\mu_1}-frac\lambda_3\mu_2}{\lambda_2}$

Formulas for the Priority Queuing Model in a Non-preemptive Queueing System

There are n priority classes with arrival rates: λ_i. And the utilization is calculated as:

$\rho_i = \frac{\lambda_i}{\mu_i}$

Using Erlang's delay formula

W_{qk}: Expected steady state time in the system spent by a type k customer.

$W_{qk} = \frac{\sum_k^n \lambda_k E(S_k^2)/2}{2(1-a_{k-1})(1-a_k)}$

Where $a_0 = 0$:

$a_k = \sum_{(} i = 1)^k \rho_i$

APPENDIX B

ADDITIONAL DATA ANALYSIS ON INPATIENT CARE

We provide additional data analysis on inpatient care that might be helpful for understanding the system dynamics.

Analyzing the Major Diagnostic Categories (MDCs)

We have analyzed the average LOS and volume for each MDC in Figure B.1, the MDCs are presented in descending order in terms of the volume presented. For example, the patients from MDC 5 have the highest volume presented and even if patients from MDC 18 do not present such a high volume, because of their long LOS, their impact on hospital occupancy is higher.

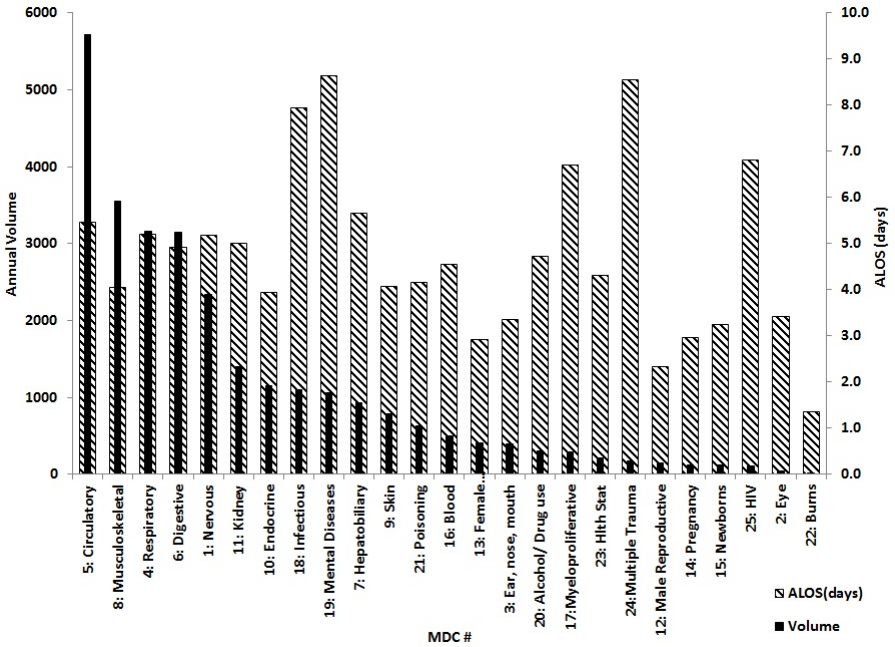


Figure B.1: ALOS and volume for each MDC

Arrival Rates

To develop insights on the data-set that is used for sampling in C#, we provide examples for arrival rates of different patient sources.

ED arrival rate

We compare the arrivals from ED observed over hours of the day with the Poisson distribution generated using the empirical means (Figure B.2). It is clear that Poisson is a good fit and the arrival rates for ED patients are often characterized with a Poisson distribution in the literature as well (McCarthy et al. [2008], Ozcan [2005]). Using data from Baystate Medical Center, Kim [2013] shows that the inter-arrival

rates for ED patients by hour of the day follow an exponential distribution, resulting in an arrival rate from ED to follow a Poisson distribution.

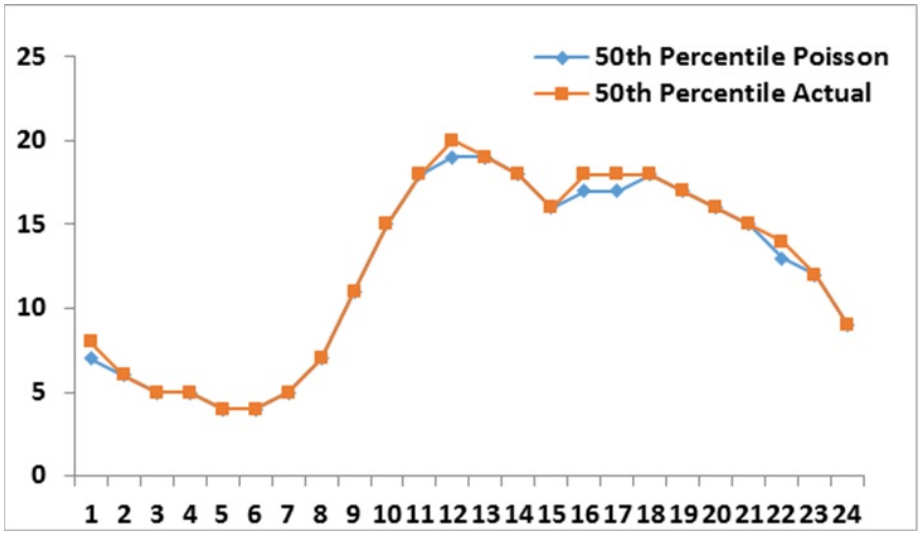


Figure B.2: Arrivals by hour – 50th percentile

Figure B.3 depicts the annual arrival rate of ED patients over hours of the day and each day of week:

As can be seen, the time-varying arrival patterns of the patients follow a similar pattern each day of the week, with peaks around the same hours of the day.

Arrival rate of elective surgeries

The annual patient volume observed for the elective patients over hours of the day and days of the week is presented in Figure B.4.

As can be seen there is virtually no demand observed on Saturday or Sunday. Thus unlike ED patients fitting the overall weekly demand to a distribution will not

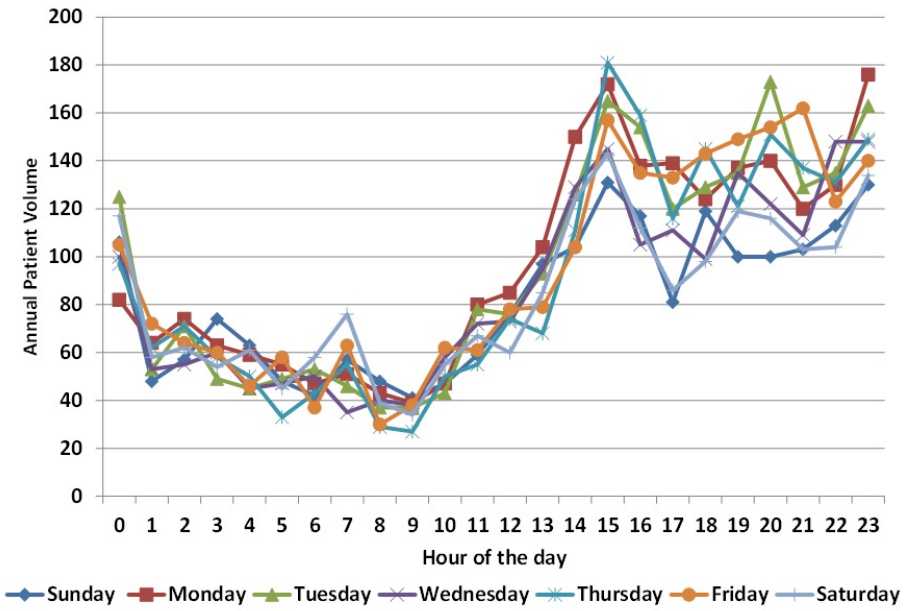


Figure B.3: Arrival rate of ED patients on each DOW and hours of the day

be appropriate, instead sampling from the data-set keeping day of week and time of day effects is more suitable.

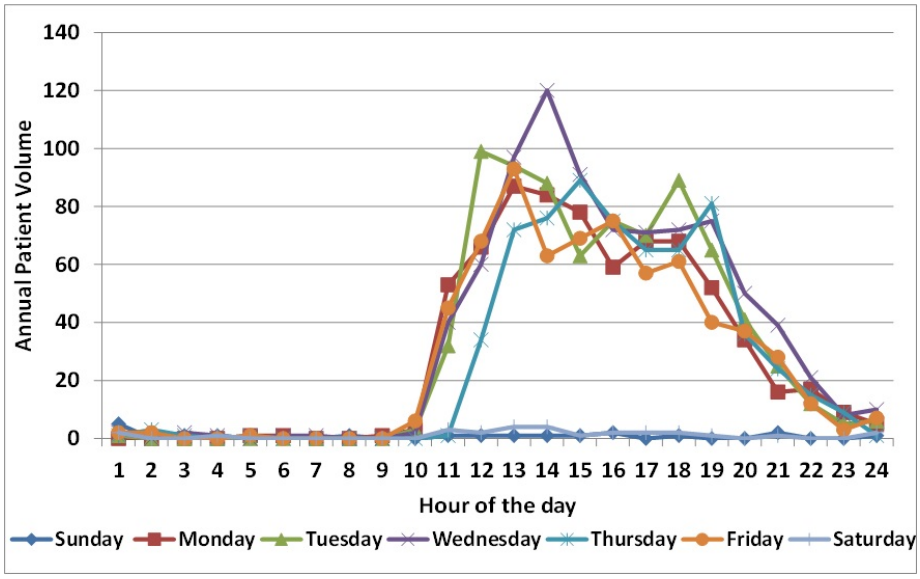


Figure B.4: Arrival rate of elective surgeries

LOS Values

To develop intuition we provide examples on the LOS patterns of two different MDCs. For one of them, lognormal is a good fit, whereas for the other MDC instead of a lognormal, beta distribution works the best. Kim [2013] use lognormal distribution in order to represent the LOS of patients for the non-ED patients admitted to Baystate Medical Center. Lognormal is considered to be a good fit for LOS durations for inpatients in the literature as well (Marazzi et al. [1998], Faddy et al. [2009]). In our Arena model, we typically use a lognormal distribution as well for characterizing the LOS of inpatients, however, C# model is developed based on sampling from the empirical values.

LOS for patients from MDC 4

For patients from MDC 4, lognormal is a good fit with the expression: -18 + LOGN(136, 114), with a square error of 0.0027 (as can be observed in Figure B.5).

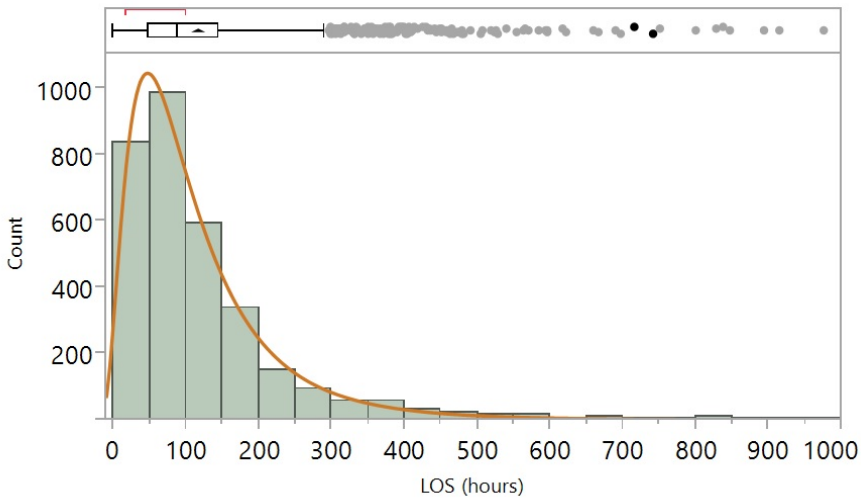


Figure B.5: Distribution of the LOS (hours) for patients from MDC 4

LOS for patients from MDC 22

Lognormal is not a good fit for patients from MDC 22 as can be observed from Figure B.6. One of the reasons is that the sample size is extremely low for this specific MDC. Indeed a beta distribution fits better with the expression: 5+49*BETA(0.938,0.696) and with a square error of 0.02.

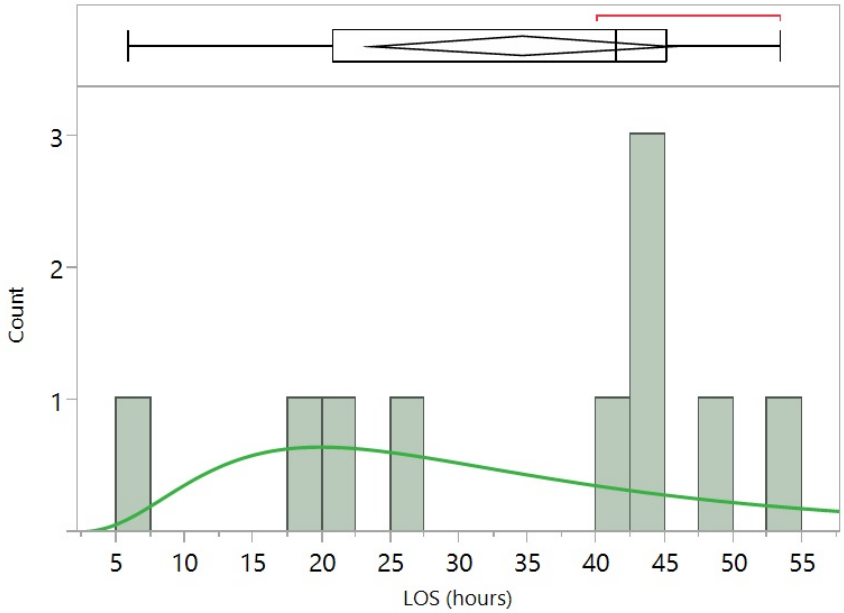


Figure B.6: Distribution of the LOS for patients from MDC 22

Elective Surgery Patients

We have started our analysis on elective surgeries by studying how the LOS and volume of scheduled surgeries vary across the days of the week (Figure B.7). As in many hospitals, there are virtually no elective surgeries performed on the weekends. The highest number of surgeries are performed on Wednesdays, and the highest LOS results from surgeries performed on Fridays (apart from the weekend surgeries which have a very small sample size).

To develop further understanding on elective surgery patterns, we have analyzed the distribution of surgeries over days of the week based on their APR-DRG severity of illness. Table B.1 indicates that the majority of the most critical surgeries are performed on Wednesdays and the least number of these surgeries happen on Thurs-

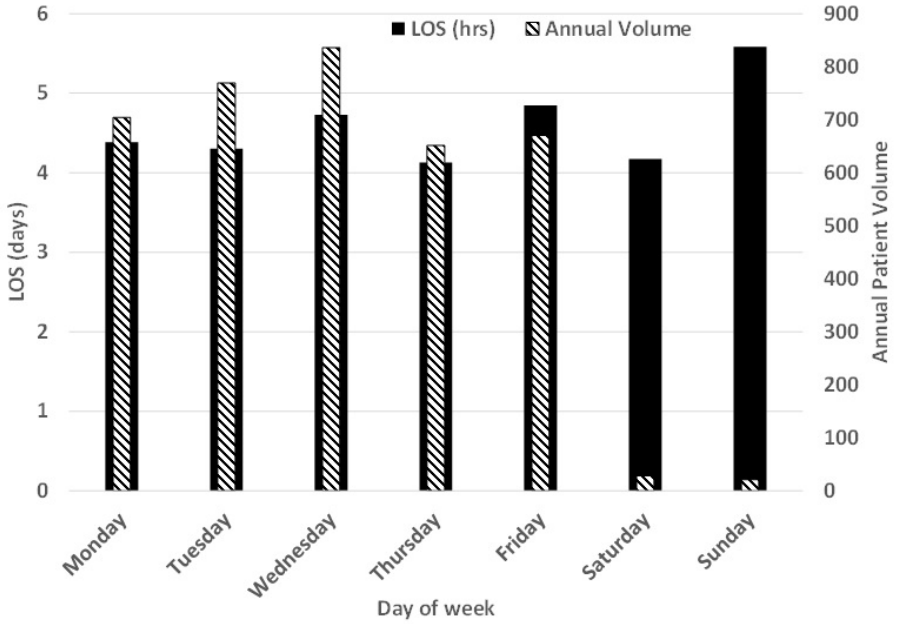


Figure B.7: LOS and volume of elective patients presented over days of the week

days. One of the main reasons is that Thursdays have the least number of surgeries performed, leading to a smaller number in critical surgeries as well.

Table B.1: Percentage of "APR-DRG Severity of Illness" categories by days of the week

	Monday	**Tuesday**	**Wednesday**	**Thursday**	**Friday**	**Saturday**	**Sunday**
Extreme	**0.18**	**0.23**	**0.24**	**0.14**	**0.16**	**0.01**	**0.03**
Major	**0.16**	**0.21**	**0.23**	**0.19**	**0.19**	**0.01**	**0.01**
Moderate	**0.2**	**0.2**	**0.25**	**0.17**	**0.17**	**0.01**	**0**
Minor	**0.19**	**0.22**	**0.19**	**0.19**	**0.19**	**0.01**	**0**

APPENDIX C

ALGORITHM OF THE INPATIENT FLOW SIMULATION MODEL

Logic Behind the C# Simulation Model

Let's say the simulation is at week 1, day 1, hour 8. The simulation looks at the historical data for Mondays 8 AM to sample the number of patients from each admission source (elective surgeries, direct admits...) with a certain MDC (or Daystay or OBS). This is used to find the number of patients to sample from that MDC in order to assign the LOS and admit unit value. Based on the admit unit, if there is a free bed, the patient is assigned a bed, if not the patient joins the queue for that unit. If the patient needs critical care, they will first visit CVICU, ICU or PICU (which is based on a discrete probability distribution derived separately for each MDC) and spend an average amount of time for that specific MDC and finally get transferred to their discharge unit. Next cycle (hour), before admitting new patients, the enqueued patients are assigned a bed. The patients are discharged after their

LOS is completed, on a FCFS basis. The bed is free after the bed turnover time is completed.

Pseudocode for the C# Simulation Algorithm

```
for w = 1..52 {For each week in the year} do
  for dow = 1...7 {For each day of the week} do
    for t = 1...24 {For each hour of the day} do
      for i = 1...8 {For each patient admission source} do
        Sample number of patients λ_{i,t}
        Sample MDC of patients M = 1..25
      end for
      for M = 1...25 {For each MDC type} do
        Sample Critical care patients and assign LOS value (Critical_LOS_M) and
        a critical admit unit specific for each MDC.
        Sample admit units j = 1..24 (A patient can either be directly admitted
        to their intercare bed or if they are critical patients they will first visit
        ICU before stepping down to these units).
        Total bed requests for unit j in time t is λ'_{j,t}.
        Sample LOS to be spent in unit j LOS_j.
      end for
      for j = 1...24 {Each admit unit} do
        for Each patient in the admitted beds do
          if TNOW is within discharge window then
```

if $TNOW \geq Dtime$ {If LOS is complete (in FCFS basis)} **then**

if $Discharged_t \leq D_t$ {The number of discharges that hour is less than the discharge capacity} **then**

$Discharged_t{+}{+}$ {Increase the number of discharged patient for that hour}

$B_t{+}{+}$ {Increase the number of available beds after bed turnover time}

end if

end if

end if

end for

for $AU = 1...3$ {Each critical admit unit} **do**

if $TNOW \geq ICU_Dtime$ {If LOS in critical unit is complete} **then**

$B_t + +$ {Increase the number of available beds after bed turnover time}

end if

end for

for Each patient requesting a bed or in the queue (in a FCFS basis) **do**

if $B_t \geq 1$ {If there is an available bed} **then**

if Patient is a critical patient **then**

$ICU_Dtime = TNOW + Critical_LOS_M$ {Admit patients and assign discharge time based on the LOS specific for each MDC category}

$B_t - -$ {Decrease the number of available beds}

$Q_t - -$ {Decrease the queue size if there is queue}

if $B_t \leq 0$ {If there is no available bed} **then**

$Q_t + +$ {Increase the queue size if there is queue}

end if

end if

if Patient is not a critical patient or is already discharged from the ICU **then**

$Dtime = TNOW + LOS$ {Admit patients and assign discharge time}

$B_t - -$ {Decrease the number of available beds}

$Q_t - -$ {Decrease the queue size if there is queue}

end if

if $B_t \leq 0$ {If there is no available bed} **then**

$Q_t + +$ {Increase the queue size if there is queue}

end if

end if

end for

end for

end for

end for

end for

Algorithm for Prioritized Discharges

To incorporate prioritization of discharges, we have only altered specific parts of the discharge algorithm as follows:

Algorithm 1 Change in the discharge algorithm for prioritized discharges

```
if TNOW is within discharge window then
  for j = 1...24 {Each admit unit} do
    if Q_t >= α {Number of patients waiting is greater than α} then
      Prioritize patients that have a high number of patients waiting to be admitted
      for Each patient in prioritized units do
        if TNOW ≥ Dtime {If LOS is complete} then
          if Discharged_t ≤ D_t {The number of discharges that hour is less than the discharge capacity} then
            Discharged_t++ {Increase the number of discharged patient for that hour}
            B_t + + {Increase the number of available beds after bed turnover time}
          end if
        end if
      end for
    end if
    if Discharged_t ≤ D_t {If there is still capacity left after discharging patients with long queues} then
      for Each patient in un-prioritized units do
        if TNOW ≥ Dtime {If LOS is complete} then
          Discharged_t + + {Increase the number of discharged patient for that hour}
          B_t++ {Increase the number of available beds after bed turnover time}
        end if
      end for
    end if
  end for
end if
```

Algorithm for Early Discharge Policy

We have tried various approaches to improve the bed congestions in Baystate Medical Center. Our initial approach was based on performing early discharges (note that this is not the same policy as early-in-the-day discharge policy–EITD). One of the main motivation for this proposed early discharge system is that, the LOS of a patient typically involves some non-value added time, due to delays. Thus, the LOS values we sample from the data-set already include some non-value added times. The main idea is to align the discharges and the admit times by pushing some of the evening discharges to the mornings as a result of discharging a subset of patients earlier than their original discharge time, as the morning are a low time for discharges.

The algorithm behind this early discharge policy is as follows: Between 9 AM 3 PM if a unit's utilization is over 85%, and if the patients have less than 6 hours to be discharged, and lastly if the truncated LOS of the patient is still more than the geometric mean for that specific MDC, then they become candidates for an early-discharge.

Making early discharges was justified by the fact that only the non-value added durations were truncated for a small subset of patients, and it did not decrease the average LOS across the inpatients significantly. Decreasing the LOS of each patient by 6 hours while keeping the LOS greater than the geometric mean, has led to significant improvements in waiting times. We have interacted with clinicians to see what level they would be comfortable with applying. However, the truncated LOS values can lead to hasty discharges, which may result in readmissions. That is why this approach was deemed infeasible, and we turned our focus to developing realistic

discharge policies which would not lead to a worse clinical outcome for patients (like readmissions).

APPENDIX D

ARENA MODEL

Why Did We Model in C# Instead of Arena?

Modeling in Arena and C#, both have their own benefits and drawbacks. Arena enables a better visualization of the model compared to C#, which can be helpful in presentations for stakeholders. On the other hand, being able to tailor the C# code instead of modeling with the "black-box" of Arena allowed us to include all the details of the complicated patient flow. In terms of computational time, Arena is a lot faster (each replication takes around 5 minutes) compared to C# model (each replication takes around 2.5 hours). One of the reasons is that we are using fitted distributions in Arena instead of sampling from the historical data. Sampling from historical data instead of using distributions allows us to more easily keep the time of day and day of week effects. Also, another important consideration is the cost factor, Arena is an expensive software for companies to invest in. On the other hand, coding with C# is more burdensome so it requires more labor, increasing expenses. For a practice implementation, Arena tool might be ideal, however, for research purposes C# offers more flexibility.

Logic Behind the Arena Simulation Model

Our analysis presented in Chapter 4 is based on our C# model. However, later we have developed an Arena model that mimics the C# code, for internal validation and to be possibly used in our partner hospital. The logic of the Arena model is explained below:

First we create separate entities for inpatients, observation and day-stay patients and assign the related attributes:

- Assign MDC for each patient using the discrete probability distributions based on the empirical proportions.

- Assign the admit unit based on the MDC, again by using a discrete probability function using the empirical data.

- Assign the LOS as a function of the admit unit and MDC, by using the best probability distribution fitted to the empirical data.

- Assign the probability that a patient requires critical care based on the MDC of a patient.

- Assign a deterministic amount of time spent in critical care based on the MDC of a patient.

Once the LOS of a patient is completed, the patient releases the bed resource and joins the discharge queue, in which the resource is a hospitalist. The number of hospitalists every hour is restricted depending on the discharge profile used. The service time is 1 hour, so the number of hospitalist implies how many discharges are allowed every hour. The patients release the hospitalist once the service time for the discharge is over.

- If the empirical discharge distribution is used, the number of hospitalists is restricted to be less than empirically observed capacities.

- If we are using the EITD policy, the number of hospitalists is practically infinite.

The hospitalists perform their discharges depending whether or not we employ a prioritization scheme.

- In the non-prioritized discharge policy, the patients are served on a FCFS basis.

- In the prioritized discharge policy, we assign a priority attribute for units that have a queue size greater than 2 (assign a lower number value for the attribute compared to the units with no queues or a lower queue size). And the patients are served based on a lowest attribute first basis in the discharge process. So the patients from the units with highest admission queues will be given priority when assigning the restricted discharge capacity.

The illustration of the Arena model with the prioritized discharge policy is shown in Figure D.1.

The results of the 10 replications is presented in Table D.1. Even though, the results do not match precisely with the outputs of the C# model, they follow the same trend as the C# results. The prioritized discharge policy performs the best, expanded discharge policy performs better than EITD, but both EITD and expanded discharge window performs better compared to the C# results. One of the main reasons for the difference in outputs is that we are sampling from historical data in C#, whereas we are using distributions fitted to the empirical data in Arena model. However, as it is discussed in Section 4.5.1 our main motivation of these runs is to

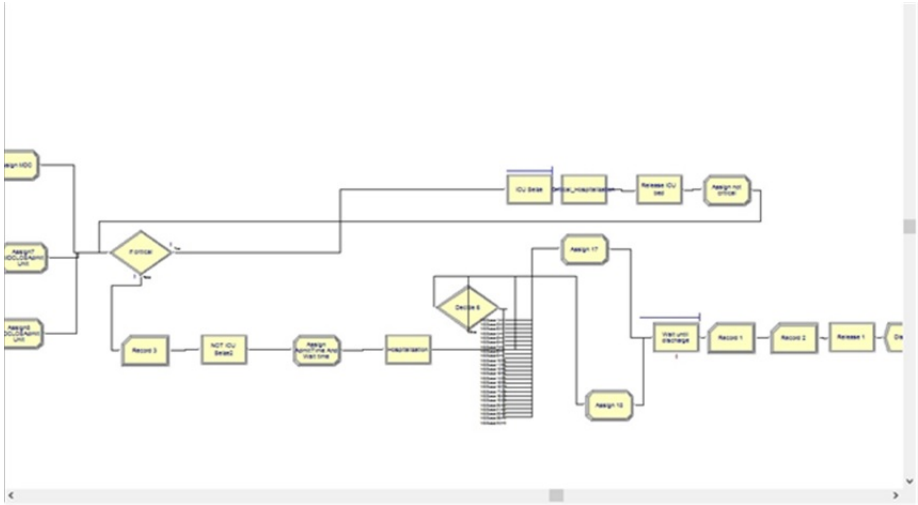


Figure D.1: Illustration of Arena model

be able to compare and analyze the impact of using different discharge policies, and Arena allows us to do that as well.

Table D.1: Results of the Arena model

Unit	10AM-7PM				10AM-9PM
	Empirical	**Max 10**	**EITD**	**Priority**	**Max 10**
S1500	0.02	0.02	0.01	0.03	0
S2	20.91	14.41	11.37	6.74	11.27
W3	4.57	5.07	3.98	2.56	3.25
S1	3.5	3.45	2.64	3.08	2.31
D6b	1.3	1.07	0.87	0.83	0.7
D6a	2.69	2.9	1.91	1.48	1.72
S3 Onc	2.08	2.19	2.09	1.9	1.92
Adolescents	0	0	0	0	0
InCh (Infants and Childrens)	0	0	0	0	0
PICU	0.33	0.34	0.33	0.29	0.3
ICU	0.02	0.02	0.02	0.02	0.02
CVICU	0	0	0	0.01	0
APTU	2.1	1.74	1.63	1.77	1.28
S4	3.32	3.68	2.31	2.24	2.08
S5	2.16	2.26	1.86	1.25	1.66
S3 MED	0.53	0.64	0.46	0.27	0.34
D5a	5.86	9.8	3.81	4.08	2.81
W4	2.24	2.6	1.93	1.26	1.46
S6400	1.1	1.59	0.72	0.72	0.41
ED	0	0	0	0	0
Surge Area	0	0	0	0	0
PACU	0	0	0	0	0
sum	**52.74**	**51.79**	**35.93**	**28.53**	**31.51**
% improvement	**NA**	**2%**	**32%**	**46%**	**40%**

APPENDIX E

INITIAL ANALYSIS ON ONGOING WORK IN INPATIENT CARE

Initial Results on Limited Discharge Capacity for Each Unit

As explained in Section 4.9, limiting the number of discharges from each admit unit to a certain threshold every hour (2 in this case) would result in a more realistic model, since data analysis has pointed out that there are at most two discharges that can happen in any unit any hour. We present the results in Table E.1 for one replication, when we incorporate unit-level constraints to the simulation model. Our preliminary analysis has shown that this does not significantly or statistically impact the queue sizes, when discharges are prioritized with these unit level constraints in mind.

Table E.1: Impact of restricting the number of discharges on queue size, where Baseline: 10 AM to 7 PM empirical discharge distribution, DP1: 10 AM-7 PM max 10, DP2: 10 AM-7 PM EITD, DP3: 10 AM-9 PM max 10, DP4: 10 AM-11 PM max 10, DP5: 10 AM-11 PM EITD, DP6: 10 AM-7 PM Empirical Priority, DP7: 10 AM-7 PM max 10 Priority, DP8: 10 AM-9 PM max 10 Priority, DP9: 24 hour discharge, DP10: 10 AM-7 PM Empirical Priority with restricted unit level discharge

Admit unit	**Baseline**	**D1**	**D2**	**D3**	**D4**	**D5**	**D6**	**D7**	**D8**	**D9**	**D10**
S2	16.46	16.31	16.36	14.81	13.37	13.08	9.75	9.78	8.98	8.01	10.19
W3	4.63	4.44	4.43	3.86	3.41	3.42	2.95	3.00	2.53	2.26	4.12
S1	3.81	3.69	3.67	3.37	3.08	3.02	1.77	1.81	1.67	2.15	0.68
D6b	1.34	1.29	1.20	1.07	0.93	0.90	0.77	0.79	0.66	0.58	0.88
D6a	0.16	0.15	0.14	0.12	0.10	0.10	0.16	0.16	0.13	0.05	0.28
S3 Onc	0.67	0.65	0.60	0.59	0.54	0.51	0.56	0.58	0.53	0.39	0.19
Adolescents	3.74	3.56	3.20	3.15	2.77	2.64	2.23	2.32	2.06	1.84	1.12
PICU	0.11	0.10	0.09	0.09	0.07	0.07	0.07	0.07	0.06	0.06	0.01
APTU	2.00	2.01	1.81	1.85	1.74	1.64	1.95	2.01	1.87	1.19	1.57
S4	2.86	2.88	2.20	2.35	1.99	1.68	1.73	1.93	1.59	1.18	1.80
S5	0.46	0.46	0.33	0.36	0.30	0.26	0.12	0.14	0.12	0.17	0.11
S3 MED	1.80	1.83	1.41	1.55	1.35	1.12	1.21	1.34	1.15	0.72	3.90
D5a	4.40	4.62	2.27	3.00	2.18	1.56	1.66	2.13	1.53	1.00	4.26
W4	3.36	4.06	1.82	2.62	1.85	1.30	2.02	2.44	1.76	0.78	2.20
SUM	**45.80**	**46.05**	**39.54**	**38.80**	**33.70**	**31.29**	**26.94**	**28.50**	**24.63**	**20.38**	**31.30**
% improvement	**NA**	**-1%**	**14%**	**15%**	**26%**	**32%**	**41%**	**38%**	**46%**	**55%**	**32%**

Initial Results on Transfers

We incorporate overflow transfers to our simulation (as explained in Section 4.10), and Table E.2 represents the results for one replication. These results indicate that the queue sizes are reduced significantly when we perform overflow transfers for units with a queue size greater than 2. We present the results for the case with transfers without the prioritized discharges (DP10) and the case both with transfers and prioritization (DP11). The improvement from these policies is extremely substantial.

Table E.2: Impact of transfers on queue size, where Baseline: 10 AM to 7 PM empirical discharge distribution, DP1: 10 AM-7 PM max 10, DP2: 10 AM-7 PM EITD, DP3: 10 AM-9 PM max 10, DP4: 10 AM-11 PM max 10, DP5: 10 AM-11 PM EITD, DP6: 10 AM-7 PM Empirical Priority, DP7: 10 AM-7 PM max 10 Priority, DP8: 10 AM-9 PM max 10 Priority, DP9: 24 hour discharge, DP10: 10 AM-7 PM Transfers without prioritization, DP11: 10 AM-7 PM Transfers with prioritization

Admit Unit	**Baseline**	**DP1**	**DP2**	**DP3**	**DP4**	**DP5**	**DP6**	**DP7**	**DP8**	**DP9**	**DP10**	**DP11**
S2	17.06	16.85	16.76	14.87	13.75	13.57	9.90	9.73	8.67	7.36	2.84	0.40
W3	5.24	5.28	5.24	4.35	3.97	4.01	4.21	4.14	3.63	2.77	1.84	0.32
S1	1.38	1.41	1.35	1.11	1.04	1.03	0.70	0.69	0.66	0.67	0.52	0.10
D6b	1.21	1.35	1.18	0.97	0.81	0.80	0.88	0.90	0.72	0.53	0.91	0.77
D6a	0.24	0.25	0.22	0.20	0.15	0.15	0.26	0.27	0.22	0.08	0.31	0.55
S3 Onc	0.21	0.21	0.19	0.19	0.17	0.16	0.19	0.20	0.17	0.12	0.19	0.07
Adolescents	1.32	1.38	1.23	1.17	1.08	1.05	1.11	1.14	0.99	0.81	1.15	0.27
APTU	1.56	1.57	1.46	1.41	1.27	1.26	1.57	1.61	1.44	1.00	1.60	1.58
S4	2.58	2.56	2.12	2.08	1.72	1.56	1.78	1.92	1.50	0.99	1.23	0.22
S5	0.29	0.29	0.22	0.23	0.20	0.18	0.11	0.12	0.10	0.13	0.16	0.25
S3 MED	5.49	5.33	4.45	4.57	4.09	3.63	3.72	4.00	3.49	2.51	1.49	0.42
D5a	10.16	8.32	5.16	5.87	4.81	3.85	3.60	4.34	3.30	2.50	1.22	0.43
W4	2.98	2.66	1.55	2.15	1.53	1.05	1.69	2.16	1.55	0.67	0.88	0.16
SUM	49.73	47.45	41.12	39.16	34.57	32.31	29.72	31.19	26.44	20.13	14.37	5.55
% improvement	NA	5%	17%	21%	30%	35%	40%	37%	47%	60%	71%	89%

APPENDIX F

ADDITIONAL DATA ANALYSIS ON SURGICAL CARE

Patient Characteristics

We use data from Mayo Clinic spine surgery practice, Rochester MN. Spine surgery related data involves 2 main OR rooms with 6 surgeons who have performed more than 2500 number of surgeries over a 5 years horizon over the years 2005 to 2011. Data available has patient-related, surgery-related and financial information on a very detailed level. We use these data properties in order to better predict and model the surgery time that enabled us to create an accurate simulation model that mimics the OR flow.

The following table summarizes the patient characteristics, their overall proportion and corresponding average surgical durations with the standard deviations.

Clinical characteristics that have a high impact on surgical duration, become vital in our categorization scheme that we developed using classification and regression tree analysis, explained in Section 5.4.1.1.

Table F.1: Patient characteristics

Characteristics	**Number of patients**	**Mean and standard deviation (or 95CI)**	**Average surgical duration ± standard deviation**
Gender	2578		
Female	1182	45.85%	4.57±2.55
Male	1396	54.15%	4.60±2.66
Age	2,578	57.5±16.4	4.58±2.61
Geographical location	2,578		
Within 5 state	2,232	86.60%	4.50±2.54
Outside 5 state	346	13.40%	5.16±2.93

Table F.2: Clinical characteristics

Characteristics	**Number of patients**	**Mean and standard deviation (or 95CI)**	**Average surgical duration ± standard deviation**
Fusion	2,468		
No	1,208	47%	3.53±2.3
Yes	1,369	53%	5.52±2.50
Number of levels	2,556	2.79±2.60 (0-9)	
Deformity	2,578		
No	2,417	93.80%	4.42±2.52
Yes	161	6.20%	7.03±2.75
Approach	2,556		
Posterior	1,842	71.95%	4.37±2.45
Lateral	112	4.38%	5.45±2.51
Anterior	498	19.38%	4.45±2.48
Staged	106	4.14%	8.18±3.11
Decompression	2,578		
No	1,033	40%	3.88±2.43
Yes	1,545	60%	5.06±2.62
Grafting	2,578		
No	1,134	43.99%	3.41±2.16
Yes	1,444	56.01%	5.51±2.56

Time-stamps

This section describes the time-stamps that were an important part of the simulation model used for outcomes projection.

Beginning of day

The beginning of day duration represents the delay from 7 AM to the OR-enter time of the patient. Days in OR are set to start at 7 AM so that the patients can be prepared for the surgery before the incision happens. However, our analysis has shown that the days typically do not start at 7 AM, as can be seen from the distribution in Figure F.1. The second peak in the graph is caused by Mondays, which start late in Mayo Clinic due to surgical fellow training objectives.

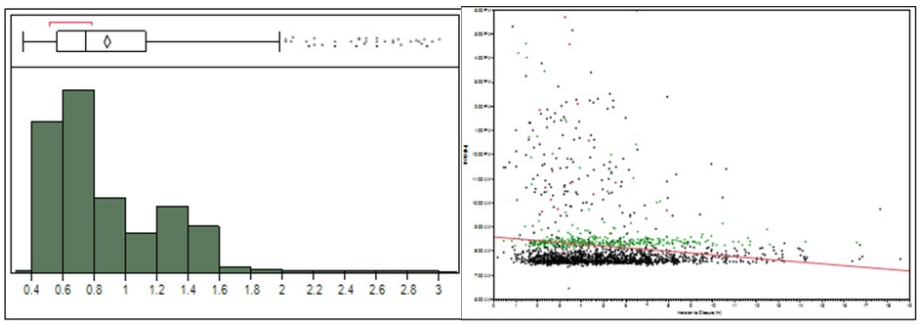


Figure F.1: The BOD distribution

Beginning of day distributions are also highly influenced by the surgery durations, as it is indicated by the figure on the right. For the first cases of the day, the longer the surgery, the earlier the surgery starts. It is characterized with a negative slope line. Thus in the simulation, different BOD distributions are used for each patient

category. The green dots on the graph indicate the start time of surgeries on late-start days (Mondays).

OR enter to incision time

Pre-incision (OR enter to incision) time includes all the time required for positioning the patient and performing anesthetic requirements. Spine surgery requires a longer pre-incision period, compared to most of the other surgeries. This is mainly because, the positioning of the patient is much more challenging than most of the other specialties. For instance, the positioning involves making sure the patients' head is at a certain angle, turning the patients while under anesthesia for a posterior surgery and so on. The surgeon is not required to be present for the preparation of the patient. Pre-incision time tends to be long even for short surgeries and almost never shorter than an hour. However, depending on the patients' clinical characteristics and the surgery type different durations are observed. Thus, there is a great variability in pre-incision times as well.

Incision to closure time

This is the actual skin to skin time, the time that the surgeon is actively performing the surgery. Because of the variety in types of surgeries performed this time distribution is highly variable as well. 10 patient groups described previously are a good proxy for incision to closure times (1 being the one with lowest and 10 being the one with highest duration).

Closure to OR exit time

This represents the time for closing the incision, again the surgeon does not necessarily have to be there. It involves closing up of the incision, which can be performed by a surgical fellow. Closure time is typically short, independent of the type of surgery performed.

OR cleaning

Even though, in literature mostly a deterministic average value is used for OR cleaning our analysis points out that the underlying distribution is highly variable (as can be seen from Figure F.2). Indeed, it is well approximated by a Normal 2 function, with a 1st peak around 40 minutes and the second one around 80 minutes.

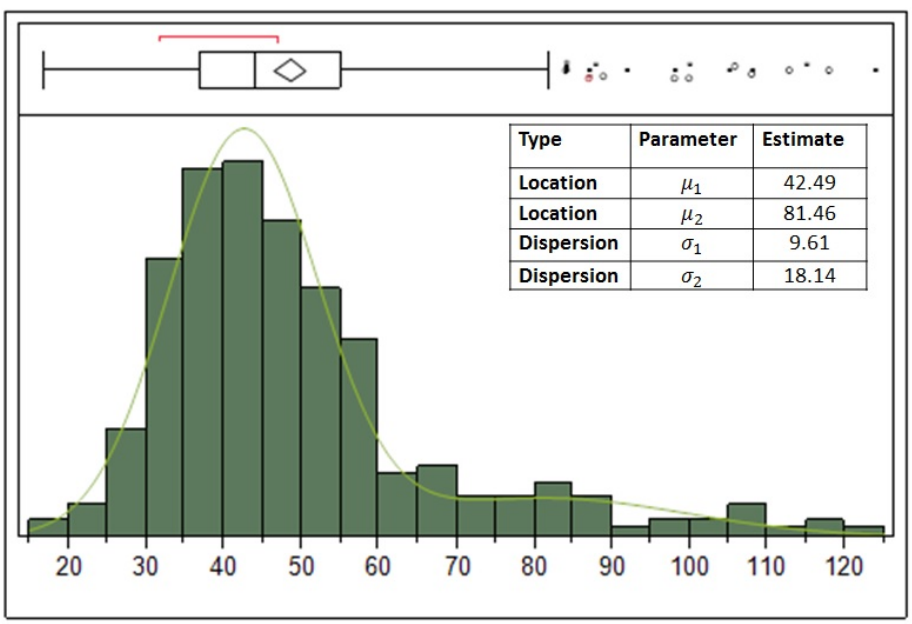


Figure F.2: OR cleaning time distribution

In order to ensure the time between surgeries was due to just OR cleaning and not surgeon related factors, we have only considered the cases for which the initial surgery's closure happens before the OR exit of the surgery in the OR room for the surgeon's next surgery. For instance we wanted to avoid taking into consideration the case presented on the left of the Figure F.3, which includes the delay from surgeon turnover time. However, the figures on the right represent the accurate calculation of OR turnover time.

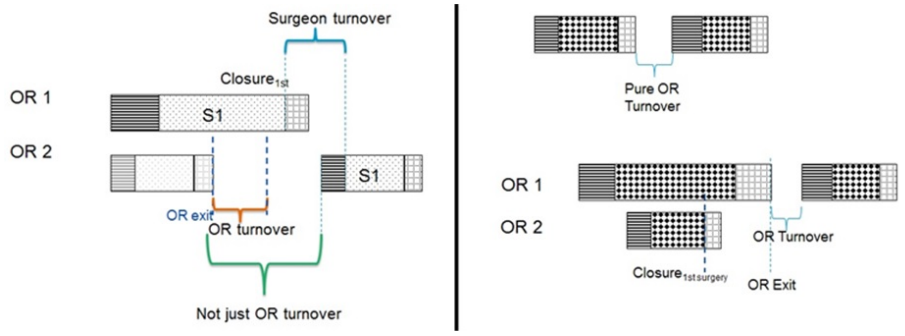


Figure F.3: Calculation of OR cleaning time

Scrub time

This time is the turnover time of a surgeon. It is most accurately calculated by the time between the points when the surgeon is done with the first surgery and the surgeon is starting the incision of the second surgery whose patient has already been prepared. Otherwise the calculation is biased, because it involves the durations other than scrubbing, like doing consults, checking up on patients and so on. Scrub time of a surgeon is generally short with a mean around 15 minutes.

APPENDIX G

SENSITIVITY ANALYSIS

We have performed sensitivity analysis through regression models, partitioning analysis and exploring the optimization framework. Regression and partitioning analysis were performed in JMP (version 9.01, SAS Institute, 2010). We derive similar conclusions from these analyses.

Partitioning Analysis

Partitioning analysis (decision tree model) allows us to examine the relationship between a response variable and multiple possible predictors. The potential predictors are evaluated using statistical methods and assessed depending on their impact on the response variable. The data is then split into two groups based on the value of the predictor (Myles et al. [2004]).

In order to understand the dynamics and the most influential factors, as a part of our sensitivity analysis, we have performed partitioning analysis. The outputs of the optimization model were used to create the experimental design. We first analyzed the most influential factors (overtime limit (o), case-mix bound width limit

(b), weight assigned to utilization ($omega$), enforced percentage of Medicare patients (m) and planning horizon(T)) for the 3 main output measures: normalized NOI, access and utilization.

Depending on the results of the initial decision tree analysis, and utilizing the ranges that the parameters were the most influential, we created experimental groups that reflect the impact of these factors on the three main outcome measures: access, utilization and NOI. We use these groups in order to conduct new experiments and run the optimization model over these specific parameter ranges and generate an unbiased data-set for the partitioning analysis.

We now summarize our findings for the key output measures:

Access analysis

Figure G.1 indicates that the most significant factor that influences access is the limit on case-mix bound width. The first split depends on whether this limit is less than or greater than 40% (Average increase in access changes from 12% to 21%). Overtime limit is the second most influential.

Thus, the impact of bound-width limit is higher than the overtime limit, in terms of improving the access. The reason is that access is directly linked to the number of surgeries, so as the surgery length decreases, more surgeries can be performed in a day. So shorter the surgeries, the higher the access. Thus, relaxing the constraint on bound-width limit allows the model to perform more of these surgeries, by allowing a flexible case-mix. Increase in overtime limit does not significantly increase the access.

This is because, longer days are not necessarily linked to more surgeries, they can be the result of very long single surgeries.

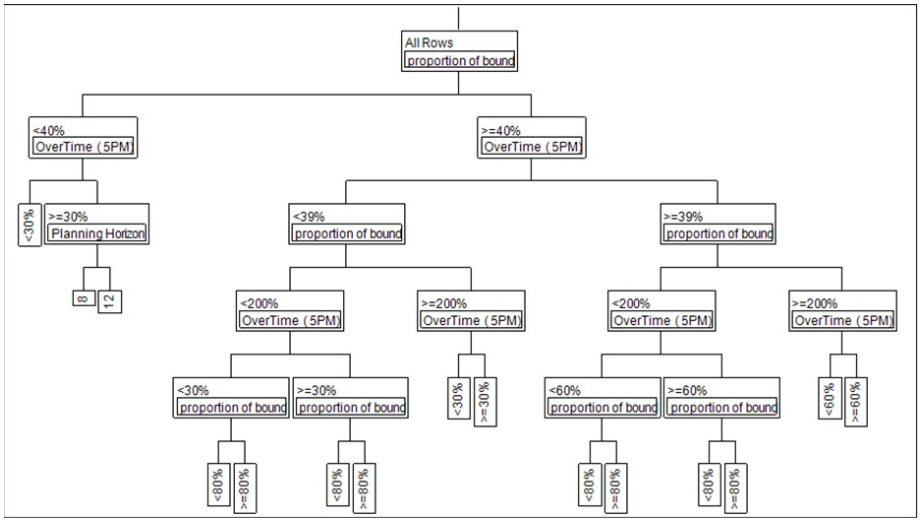


Figure G.1: Partitioning for access

NOI analysis

The primary factor that affects NOI is the overtime limit, as can be seen from Figure G.2. If the overtime limit is over some threshold (40%), the impact of case-mix bound width limit becomes significant. Surprisingly, weight assigned in the objective function does not significantly impact the normalized NOI.

Higher NOI is a result of longer surgeries, since these are the more complex patients with longer hospitalization periods. And the longer the LOS, the higher the revenue. Thus, if the days are allowed to end later, the optimization will assign longer surgeries resulting in a higher NOI. However, even if relaxing the overtime limit will improve NOI, longer surgeries (patients from higher numbered categories)

also occur more rarely than shorter surgeries. So the bound width limit is a limiting factor as well, however, has a lower impact on NOI.

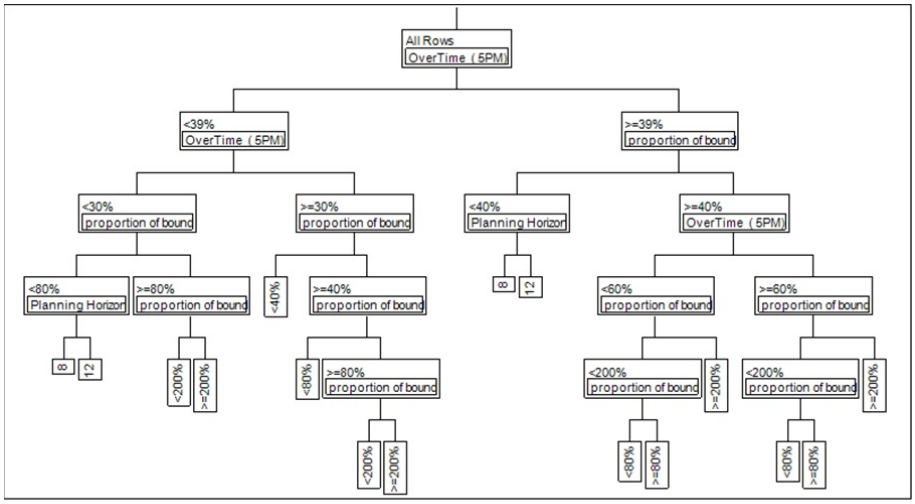


Figure G.2: Partitioning for NOI

Utilization analysis

The primary factor that affects utilization is the overtime limit (Figure G.3). The higher the overtime limit, the higher the prime-time period utilization. The utilization can be increased within a wide range of case-mix bound widths. The longer surgeries are not necessarily what drives a higher utilization, different combinations of surgery pairs can result in a high utilization as well. As long as these pairs increase the length of surgical days, they will have the same impact as long surgeries. So the bound width limit is not as restricting when trying to achieve a high utilization.

We have also observed that what drives NOI also drives utilization. However, a high utilization rate does not necessarily imply higher access, since access is primarily

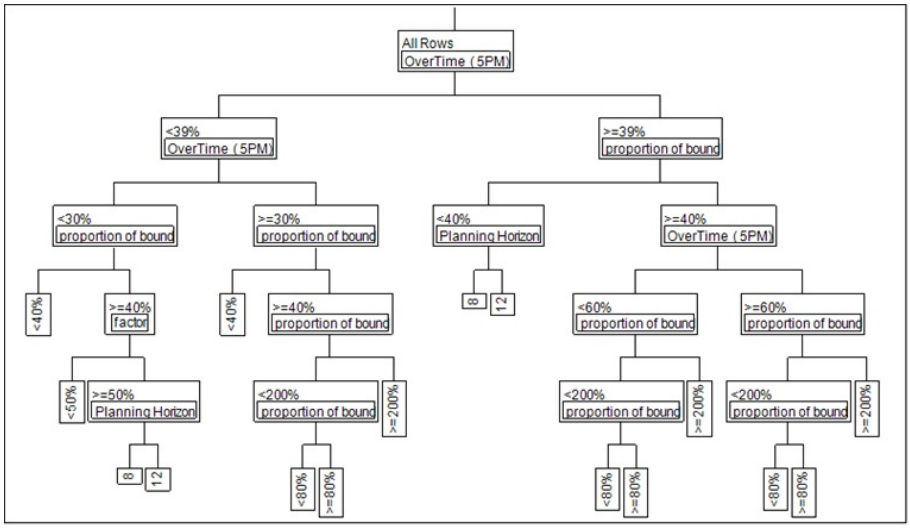


Figure G.3: Partitioning for utilization

influenced by the limit on bound width and long surgical days are not necessarily a result of a greater number of surgeries.

Regression Analysis

Regression analysis is typically used to estimate the relationship between a dependent variable (access, utilization and NOI) and the independent variables (utilization, overtime and case-mix bound width limit, enforced Medicare percentage, planning horizon). Thus, regression analysis enables us to test how the dependent variable changes when the independent variables are varied, while the other independent variables are fixed (Kleinbaum et al. [2013], Kutner [1996]). Partitioning and regression analysis led to similar conclusions in terms of the relationship among constraints and output measures.

Access analysis

The results presented in Table G.1 and Figure G.4 show that weight assigned to utilization, overtime and case-mix bound width limit has a significant impact on access (number of surgeries performed). There is a positive correlation between the percentage of overtime and case-mix bound width with the increase in access. On the other hand, planning horizon and percentage of Medicare patients do not.

Table G.1: Parameter estimates for access

Term	**Estimate**	**Std Error**	**t Ratio**	**Prob**$> \lvert t \rvert$
Intercept	0.1299109	0.007672	16.93	<.0001*
Planning Horizon	0.0001462	0.00005	2.93	0.0038*
medicare proportion	-0.023457	0.017746	-1.32	0.1879
OverTime (5PM)	0.0804239	0.008114	9.91	<.0001*
Case-mix bound width	0.0285461	0.0014	20.4	<.0001*
Weight	0.0311085	0.006511	4.78	<.0001*

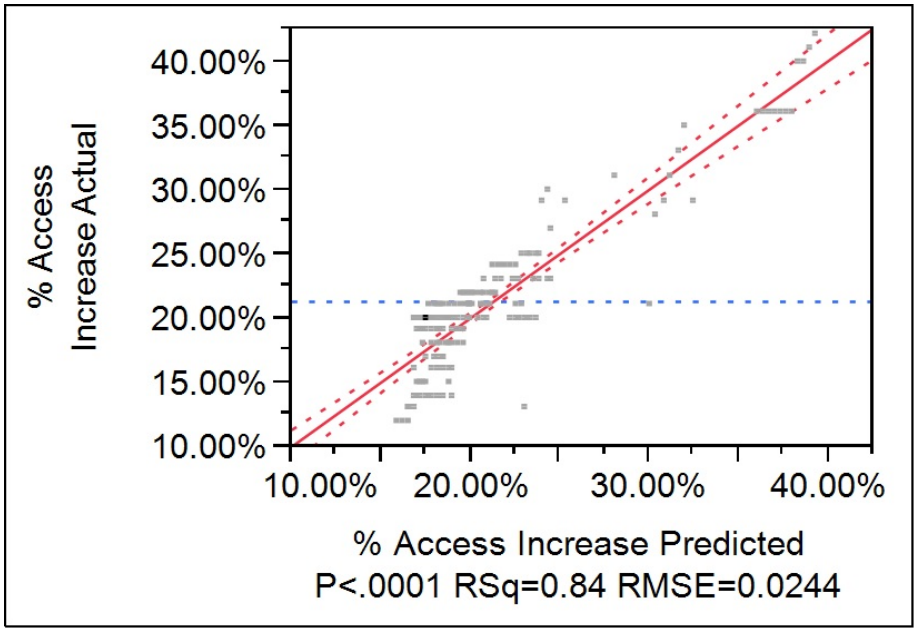


Figure G.4: Regression analysis for access

NOI analysis

Regression analysis indicates that NOI is influenced by all of the independent variables except for the planning horizon (Table G.2 and Figure G.5). Out of the three output measures only NOI is influenced by the enforced Medicare percentage, since different reimbursement policies have a significant impact on revenue.

Table G.2: Parameter estimates for NOI

Term	Estimate	Std Error	t Ratio	Prob> $\|t\|$
Intercept	0.7486846	0.022687	33	<.0001*
Planning Horizon	0.000364	0.000148	2.47	0.0146*
medicare proportion	-0.537294	0.052478	-10.24	<.0001*
OverTime (5PM)	0.1051077	0.023993	4.38	<.0001*
Case-mix bound width	0.0197918	0.004139	4.78	<.0001*
Weight	-0.156976	0.019252	-8.15	<.0001*

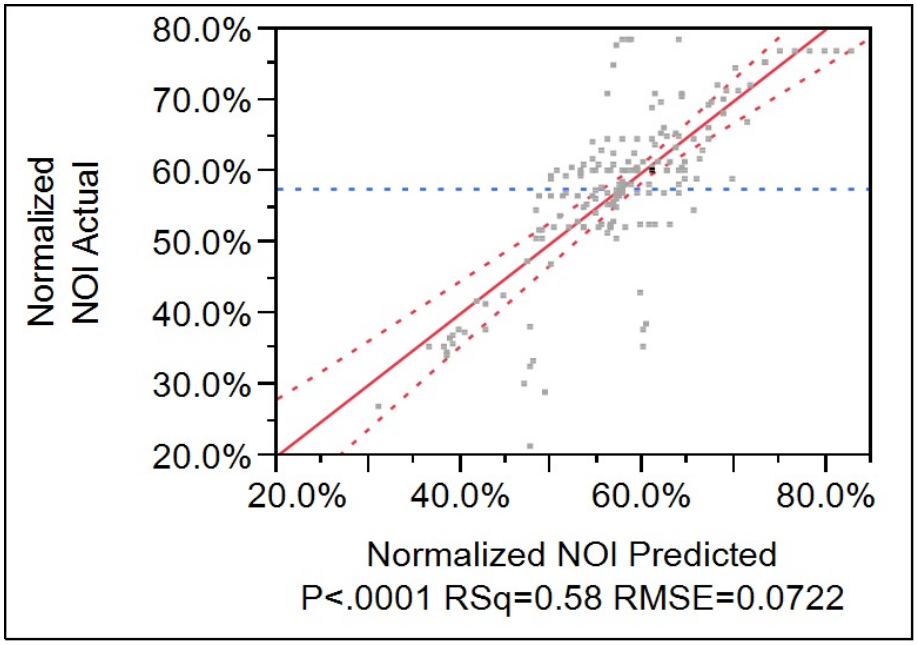


Figure G.5: Regression analysis for NOI

Utilization analysis

Table G.3 and Figure G.6 show that utilization is primarily influenced by overtime and case-mix bound width limit. Weight assigned in the objective function, planning horizon and percentage of Medicare patients do not have a significant impact on utilization.

Table G.3: Parameter estimates for utilization

Term	Estimate	Std Error	t Ratio	Prob> $\|t\|$
Intercept	0.7248899	0.011331	63.97	<.0001*
Planning Horizon	0.0001741	7.37E-05	2.36	0.0193*
medicare proportion	-0.032645	0.02621	-1.25	0.2146
OverTime (5PM)	0.0745261	0.011983	6.22	<.0001*
Case-mix bound width	0.0153224	0.002067	7.41	<.0001*
Weight	0.0259018	0.009616	2.69	0.0077*

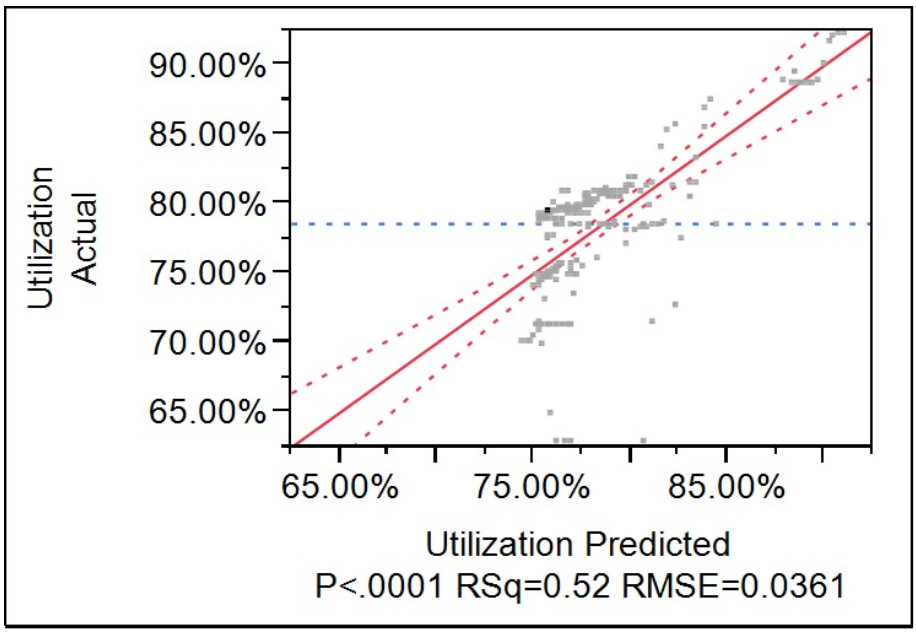


Figure G.6: Regression analysis for utilization

BIBLIOGRAPHY

A.E. Abouleish, D. S. Prough, C. W. Whitten, and M. H. Zornow. The effects of surgical case duration and type of surgery on hourly clinical productivity of anesthesiologists. *Anesth Analg*, 97, 2003.

J. Abraham and M. C. Reddy. Challenges to inter-departmental coordination of patient transfers: a workflow perspective. *International Journal of medical informatics*, 79(2):112–22, 2010.

I. Adan, J. Bekkers, N. Dellaert, J. Vissers, and X. Yu. Patient mix optimisation and stochastic resource requirements: A case study in cardiothoracic surgery planning. *Health Care Manag Sci.*, 12(2):129–41, 2009.

L. A. Aday and R. Andersen. A framework for the study of access to medical care. *Health services research*, 9(3):208, 1974.

G. C. Alexander, J. Kurlander, and M.K. Wynia. Physicians in retainer ("concierge") practice. A national survey of physician, patient, and practice characteristics. *J Gen Intern Med.*, 20(12):1079–1083, 2005.

J. Altschuler, D. Margolius, T. Bodenheimer, and K. Grumbach. Estimating a reasonable patient panel size for primary care physicians with team-based task delegation. *The Annals of Family Medicine*, 10(5):396–400, 2012.

M. Armony, S. Israelit, A. Mandelbaum, Y. Marmor, Y. Tseytlin, and G. Yom-Tov. Patient flow in hospitals: A data-based queueing perspective, Sep 2012. URL `http://www.stern.nyu.edu/om/faculty/armony`. Working paper.

ASCA. What is an ASC?, 2013. URL `http://www.ascassociation.org/ASCA/AboutUs/WhatisanASC`. Available online.

The Press Association. Unnecessary ward moves "harm elderly patients", warn academics, 2013. URL `http://www.nursingtimes.net/nursing-practice/`

clinical-zones/older-people/unnecessary-ward-moves-harm-elderly-patients-warn-academics/5061707.

H. J. Balasubramanian, R. Banerjee, B. Denton, J. Naessens, D. Wood, and J. Stahl. Improving clinical access and continuity using physician panel redesign. *Journal of General Internal Medicine*, 25(10):1109–15, 2010.

H. J. Balasubramanian, A. Muriel, and L. Wang. The impact of flexibility and capacity allocation on the performance of primary care practices. *Flexible Services and Manufacturing Journal*, 2011. ISSN 1936-6582. doi: 10.1007/s10696-011-9112-5.

H. J. Balasubramanian, A. Muriel, A. Ozen, L. Wang, J. Hippchen, and X. Gao. In Brian Denton, editor, *Springer Book on Healthcare Operations Management*, volume 184, chapter Capacity allocation and flexibility in primary care, pages 205–228. Springer, 2013.

J. Banks, J.S. Carson, B.L. Nelson, and D.M. Nicol. *Discrete-event system simulation*. Pearson Prentice Hall, Upper Saddle River, NJ, 2004.

S. Batun, B. T. Denton, T. R. Huschka, and A. J. Schaefer. Operating room pooling and parallel surgery processing under uncertainty. *INFORMS Journal on Computing*, 23(2):220–37, 2011.

R. Bekker and P. M. Koeleman. A model of ICU bumping. *Operations Research*, 14: 1564–74, 2010.

R. Bekker, J. E. Helm, and M. P. Van Oyen. Discharge planning to mitigate bed block and er overcrowding. 2014.

T. Best, B. Sandikci, D. Eisenstein, and D. Meltzer. Managing hospital bed capacity through partitioning care into focused wings. *Chicago Booth Research Paper*, 12 (64), 2012.

T. W. Bice and S. B. Boxerman. A quantitative measure of continuity of care. *Medical care*, page 347-49, 1977.

J. T. Blake and J. Donald. Mount Sinai hospital uses integer programming to allocate operating room time. *Interfaces*, 32(2):63–73, 2002.

T. Bodenheimer and H. H. Pham. Primary care: Current problems and proposed solutions. *Health Affairs*, 29(5):799–805, 2010.

E. Brockmeyer, H.L. Halstrom, and A. Jensen. The life and works of A.K. Erlang. *Transactions of the Danish Academy of Technical Science 2*, 1948.

W. Cao and G.J. Lim. An introduction to optimization models and applications in healthcare delivery systems. In Yuehwen Yih, editor, *Handbook of Healthcare Delivery Systems*. CRC Press, January 2011.

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

S. Carr and S.D. Roberts. In Yuehwen Yih, editor, *Handbook of Healthcare Delivery Systems*, chapter 14: Computer Simulation in Healthcare. CRC Press, January 2011.

CDC. Rising health care costs are unsustainable, 2011. URL `http://www.cdc.gov/workplacehealthpromotion/businesscase/reasons/rising.html`.

S. Chakraborty, K. Muthuraman, and M. Lawley. Sequential clinical scheduling with patient no-shows and general service time distributions. *IIE Transactions*, 42(5): 354–366, 2010.

M.A. Chen, J.P. Hollenberg, W. Michelen, J.C. Peterson, and L.P. Casalino. Patient care outside of office visits: A primary care physician time study. *J Gen Intern Med.*, 26(1):58–63, 2011.

S. Choi and W. E. Wilhelm. An analysis of sequencing surgeries with durations that follow the lognormal, gamma, or normal distribution paper. *IIE Transactions on Healthcare Systems Engineering*, 2012.

CMS. Are You a Hospital Inpatient or Outpatient?, February 2011. URL `http://www.medicare.gov/Pubs/pdf/11435.pdf`. Available online.

M. D. Cohen and B. Hilligoss. The published literature on handoffs in hospitals: deficiencies identified in an extensive review. *Quality and Safety in Health Care*, 19(6):493–7, 2010.

K. Coleman, R. J. Reid, E. Johnson, C. Hsu, T. R. Ross, P. Fishman, and E. Larson. Implications of reassigning patients for the medical home: A case study. *Annals of Family Medicine*, 8(6):493–98, 2010.

T.R. Collins. Should hospitalists be concerned about the pchm model? *The Hospitalist*, July 2012a.

T.R. Collins. Hospitalists should prepare for the patient centered medical home. *The Hospitalist*, July 2012b.

J. Colwell. Measuring hospitalist workload. *ACP Hospitalist*, 2013.

H. Davies and R. Davies. Simulating health systems = modelling problems and software solutions. *Eur. J. Oper. Res.*, 87:35–44, 1995.

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

B. T. Denton, A. J. Miller, H. J. Balasubramanian, and T. R. Huschka. Optimal allocation of surgery blocks to operating rooms under uncertainty. *Oper. Res.*, 58 (4):802–16, 2010.

F. Dexter, E.U. Dexter, and J. Ledolter. Influence of procedure classification on process variability and parameter uncertainty of surgical case durations. *Anesth Analg*, 110(4):1155–63, 2010.

DH. Achieving timely simple discharge from hospital: A toolkit for the multi-disciplinary team, 2004. URL `http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH 4088366`. Available online.

J. Donz, S. Lipsitz, D. W. Bates, and J. L. Schnipper. Causes and patterns of readmissions in patients with common comorbidities: retrospective cohort study. *BMJ: British Medical Journal*, 347, 2013.

L. N. Dyrbye and T. D. Shanafelt. Physician burnout: a potential threat to successful health care reform. *JAMA*, 305(19):2009–10, 2011.

L.N. Dyrbye, C.P. West, T.C. Burriss, and T.D. Shanafelt. Providing primary care in the United States: The work no one sees. *Arch Intern Med.*, 172(18):1420–1421, 2012.

S. A. Erdogan, B. T. Denton, J. J. Cochran, L. A. Cox, P. Keskinocak, J. P. Kharoufeh, and J. C. Smith. Surgery planning and scheduling. *Wiley Encyclopedia of operations research and management science*, 2011. URL `http://ca. wiley. com/WileyCDA/Section/id-380199. html`.

S. Espin, L. Lingard, G.R. Baker, and G. Regehr. Persistence of unsafe practice in everyday work: an exploration of organizational and psychological factors constraining safety in the operating room. *Qual Saf Health Care*, 15(3):165–170, 2006.

M. Faddy, N. Graves, and A. Pettitt. Modeling length of stay in hospital and other right skewed data: Comparison of phase-type, gamma and log-normal distributions. *Value in Health*, 12(2):309–314, 2009.

FCA. Hospital Discharge Planning: A Guide for Families and Caregivers, 2013. URL `http://www.caregiver.org/caregiver/jsp/content_node.jsp`. Available online.

H.H. Forsberg, H. Aronsson, C. Keller, and S. Lindblad. Managing health care decisions and improvement through simulation modeling. *Q Manage Health Care*, 20(1):15–20, 2011.

R. Freund. Professor George Dantzig : Linear programming founder turns 80, November 1984. URL `http://www.stanford.edu/group/SOL/dantzig.html`.

B. Fries. Bibliography of operations research in health-care systems. *Operation Research*, 24(5):801–4, 1980.

X. Gao, H. Xu, and D. Ye. Asymptotic behavior of tail density for sum of correlated lognormal variables. *International Journal of Mathematics and Mathematical Sciences*, 2009, 2009.

F.B. Gilbreth. Motion study in surgery. *Can J Med Surg*, 40(22):22–31, 1916.

J. M. Gill and A. Mainous. The role of provider continuity in preventing hospitalizations. *Archive of Family Medicine*, 7:352–357, 2010.

J. J. Glasheen, G. J. Misky, M. B. Reid, R. A. Harrison, B. Sharpe, and A. Auerbach. Career satisfaction and burnout in academic hospital medicine. *Archives of internal medicine*, 171(8):782–90, 2011.

L. V. Green. Queueing analysis in healthcare. In R. W. Hall, editor, *Patient Flow: Reducing Delay in Healthcare Delivery*, pages 281–308. Springer-Verlag, 2006.

L. V. Green. The vital role of operations analysis in improving healthcare delivery. *MSOM*, 14:488–494, 2012.

L. V. Green and V. Nyugen. Strategies for cutting hospital beds: the impact of patient service. *Health Services Research*, 36:421–42, 2001.

L. V. Green and S. Savin. Reducing delays for medical appointments: A queueing approach. *Operations Research*, 56(6):1526–38, 2008.

L. V. Green, S. Savin, and M. Murray. Providing timely access to care: What is the right patient panel size? *The Joint Commission Journal on Quality and Patient Safety*, 33:211–18, 2007a.

L. V. Green, S. Savin, and Y. Lu. Primary care physician shortages could be eliminated through use of teams, nonphysicians, and electronic communication. *Health Affairs*, 32(1):11–19, 2013.

L.V. Green. How many hospital beds? *Inquiry*, 39(4):400–12, 2003.

L.V. Green. Queueing theory and modeling. In Yuehwen Yih, editor, *Handbook of Healthcare Delivery Systems*. CRC Press, January 2011.

L.V. Green, P.J. Kolesar, and W. Whitt. Coping with time-varying demand when setting staffing requirements for a service systems. *Production and Operations Management*, 16(1):13–39, 2007b.

D. Gross and C. Harris. *Fundamentals of Queueing Theory*. John Willey and Sons, Inc., 1985.

C. Grossmann, W. A. Goolsby, L. Olsen, J. M. McGinnis, Institute of Medicine, and National Academy of Engineering. *Engineering a Learning Healthcare System:A Look at the Future: Workshop Summary*. The National Academies Press, 2011.

F. Guerriero and R. Guido. Operational research in the management of the operating theatre: a survey. *Health Care Management Science*, 14(1):89–114, 2011.

P. Guo and R. Hassin. Equilibrium and optimal strategies for placing duplicate orders in queues. Working paper, 2012.

D. Gupta. Surgical suites' operations management. *Oper. Management*, 16(6):689–700, 2007.

D. Gupta and L. Wang. Revenue management for a primary-care clinic in the presence of patient choice. *Operations Research*, 56(3):576–592, May-June 2008.

S. Gurumurthi and S. Benjaafar. Modeling and analysis of flexible queueing systems. *Naval Research Logistics (NRL)*, 51(5):755–782, 2004.

W. M. Hancock and P.F. Walter. The use of computer simulation to develop hospital systems. *ACM SIGSIM Simulation Digest*, 10(4):28–32, 1979.

J. Harding. Study of discharge communications from hospital doctors to an inner london general practice. *J Gen Intern Med*, 17(2):186–92, 2002.

G. Harris. Family physician can't give away solo practice. NY Times, 2011.

J. E. Helm and M.P. Van Oyen. Design and optimization methods for elective hospital admissions. Technical Report 11-01, University of Michigan, Industrial and Operations Engineering, February 2010.

J. E. Helm, S. AhmadBeygi, and M.P. Van Oyen. Design and analysis of hospital admission control for operational effectiveness. *POMS*, 20(3):359–74, 2011.

J.E. Helm, M. Lapp, and B.D. See. Characterizing an characterizing an effective hopsital admission scheduling and control management system: A genetic algorithm approach and control management system: A genetic algorithm approach. *Proceedings 42nd Conf on Winter Sim.*, 2010.

HHS. Shortage designation: Hpsas, muas & mups. online, 2009.

B. Hilligoss and M. D. Cohen. The unappreciated challenges of between-unit handoffs: negotiating and coordinating across boundaries. *Annals of emergency medicine*, 61(2):155–60, 2013.

HMA. To free beds for new admissions, triage best candidates for early discharge. *HcPro*, 2(10), October 2006.

A.N. Hofer, J.M. Abraham, and I. Moscovice. Expansion of coverage under the patient protection and affordable care act and primary care utilization. *The Milbank Quarterly*, 89(1):69–89, 2011.

IOM and NAE. Rising health care costs are unsustainable, 2012. URL `http://www.iom.edu/ /media/Files/Activity`.

IOM, NAE, Committee on Engineering, and the Health Care System. *Building a Better Delivery System: A New Engineering/Health Care Partnership.* The National Academies Press, 2005. URL `http://www.ncbi.nlm.nih.gov/books/NBK22832/`.

B. Jaeker, J. Alexandra, and A. L. Tucker. Hurry up and wait: Differential impacts of congestion, bottleneck pressure, and predictability on patient length of stay. *Harvard Business School*, 2012. Working paper.

J. James. Health policy brief: Medicare hospital readmissions reduction program. *Health Affairs*, 2013.

S.F. Jencks, M.V. Williams, and E.A. Coleman. Rehospitalizations among patients in the medicare fee-for-service program. *New England Journal of Medicine*, 360 (8):1418–28, 2009.

W. C. Jordan and S. C. Graves. Principles on the benefits of manufacturing process flexibility. *Management Science*, 41(4):577–94, 1995.

D. Kc and C. Terwiesch. Impact of workload on service time and patient safety: An econometric analysis of hospital operations. *Management Science*, 55(9):14861498, 2009.

B.T. Kealey and B.R. Asplin. Throughput and congestion in the modern hospital. conference proceedings. New Orleans, LA, April 2004.

C. S. Kim, W. Lovejoy, M. Paulsen, R. Chang, and S.A. Flanders. Hospitalist time usage and cyclicality: opportunities to improve efficiency. *Journal of Hospital Medicine*, 5(6):329–34, 2010.

E. S. Kim. Using computer simulation to study hospital admission and discharge processes. Master's thesis, University of Massachusetts Amherst, 2013.

D. Kleinbaum, L. Kupper, A. Nizam, and E. Rosenberg. *Applied Regression Analysis and Other Multivariable Methods*. Cengage Learning, 2013.

Jim Kling. Burnout high among hospitalists. *Arch Intern Med*, 8:782–785, 2011.

S. Knox and H. Britt. The contribution of demographic and morbidity factors to self-reported visit frequency of patients: A cross-sectional study of general practice patients in australia. *BMC Family Practice*, 5(1):17, 2004.

Kolker, A. Queuing Analytic Theory and Discrete Events Simulation for Healthcare: Right Application for the Right Problem, 2010. URL `http://www.irma-international.org/viewtitle/49971/`. Online.

B. Kollberg, J. J. Dahlgaard, and P. O. Brehmer. Measuring lean initiatives in health care services: issues and findings. *International Journal of Productivity and Performance Management*, 56(1), 2006.

R. Kopach-Konrad, M. Lawley, M. Criswell, I. Hasan, S. Chakraborty, J. Pekny, and B.N. Doebbeling. Applying systems engineering principles in improving health care delivery. *J Gen Intern Med.*, 22(3):431–37, 2007.

S. Kripalani, A. T. Jackson, J. L. Schnipper, and E. A. Coleman. Promoting effective transitions of care at hospital discharge: a review of key issues for hospitalists. *Journal of Hospital Medicine*, 2(5):314–23, 2007a.

S. Kripalani, F. LeFevre, C. O. Phillips, M. V. Williams, P. Basaviah, and D. W. Baker. Deficits in communication and information transfer between hospital based and primary care physicians. *Jama*, 297(8):831–841, 2007b.

V. Kulkarni. *Modeling and Analysis of Stochastic Systems*. Chapman Hall, CRC, 1995.

M. H. Kutner. *Applied linear statistical models*, volume 4. McGraw-yHill/Irwin, Chicago, 1996.

L. LaGanga and S. Lawrence. Clinic overbooking to improve patient access and increase provider productivity. *Decision Sciences*, 38(2), 2007.

A. Law and W. Kelton. *Simulation modeling and analysis*. McGraw-Hill, 1991.

R. Levi and A. Prestipino. Commentary-driving new science of healthcare delivery: What does it take to make an impact? *MSOM*, 2012. Invited paper to a special issue of Manufacturing & Services Operations Management (MSOM) on healthcare operations management. To Appear.

N. Liu and T. D'Aunno. The productivity and cost-efficiency of models for involving nurse practitioners in primary care: A perspective from queueing analysis. *Health services research*, 47(2):594–613, 2012.

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

N. Liu, S. Finkelstein, and L. Poghosyan. A new model for nurse practitioner utilization in primary care: Increased efficiency and implication. *Health Care Management Review (forthcoming)*, 2012.

Liu, N. and Ozen, A. and Balasubramanian, H.J. Primary Care Practice Design under Case Mix: Joint Consideration of Access to Care and Continuity of Care, 2014. Working paper.

Phyllis Maguire. What's the ideal number of patients to see? *Today's Hospitalist*, July 2009.

Phyllis Maguire. A day in the life. *Today's Hospitalist*, October 2010.

M. Mahon and J. Weymouth. U.S. spends far more for health care than 12 industrialized nations, but quality varies, May 2012. URL `http://www.commonwealthfund.org/News/News-Releases/2012/May/US-Spends-Far-More-for-Health-Care-Than-12-Industrialized-Nations-but-Quality-Varies.aspx`.

L. Manchikanti, D. Caraway, A.T. Parr, B. Fellows, and Joshua A. Hirsch. Patient protection and affordable care act of 2010: Reforming the health care reform for the new decade. *Pain Physician*, 14:35–67, 2011.

A. Marazzi, F. Paccaud, C. Ruffieux, and C. Beguin. Fitting the distributions of length of stay by parametric models. *Medical Care*, 36(6):915–927, 1998.

A.B. Martin, D. Lassman, B. Washington, A. Catlin, and National Health Expenditure Accounts Team. Growth in US health spending remained slow in 2010; health share of gross domestic product was unchanged from 2009. *Health Affairs*, 31(1): 208–19, 2012.

M. L. McCarthy, S. L. Zeger, R. Ding, D. Aronsky, N. R. Hoot, and G. D. Kelen. The challenge of predicting demand for emergency department services. *Academic Emergency Medicine*, 15(4):337–346, 2008.

Mary McMahon. What is a telemetry unit? *Conjecture Corporation*, 2014.

MDH. Health care homes-payment methodology development, 2010. URL `http://www.health.state.mn.us/healthreform/homes/payment/index.html`.

D. Mechanic, D. D. McAlpine, and M. Rosenthal. Are patients' office visits with physicians getting shorter? *New England Journal of Medicine*, 344(3):198–204, 2001.

MedPac. Promoting greater efficiency in medicare, chapter 5: Payment policy for inpatient readmissions. Technical report, Report to Congress, 2007. URL `http://www.medpac.gov/chapters/Jun07_Ch05.pdf`.

J. B. Montgomery and K. Davis. The hospital patient flow model: A simulation decision support tool. *Conference Proceedings, 2013 Society for Health Systems: Healthcare Systems Process Improvement Conference*, 2013.

M. Murray and D. M. Berwick. Advanced access: Reducing waiting and delays in primary care. *Journal of the American Medical Association*, 289(8):1035–40, 2003.

M Murray and C. Tantau. Same-day appointments: Exploding the access paradigm. *Fam Pract Manag.*, 7(8):45–50, 2000.

M. Murray, M. Davies, and B. Boushon. Panel size: How many patients can one doctor manage? *Fam Pract Manag.*, 14(4):44–51, 2007.

K. Muthuraman and M. Lawley. Stochastic overbooking model for outpatient clinical scheduling with no shows. *IIE Transactions*, 40(9), 2008.

A. J. Myles, R. N. Feudale, Y. Liu, N. A. Woody, and S. D. Brown. An introduction to decision tree modeling. *Journal of Chemometrics*, 18(6):275–285, 2004.

J. Naessens, R. Stroebel, D. Finnie, N. Shah, A. Wagie, W. Litchy, P. Killinger, T. O'Byrne, D. Wood, and R. Nesse. Effect of multiple chronic conditions among working-age adults. *American Journal of Managed Care*, 17(2):118–22, 2011.

L. Nan and L. Li. Estimating the distribution of surgical durations based on eliminating the abnormal data. *Service Systems and Service Management (ICSSSM)*, pages 1–5, 2011.

Nelhin. Discharge from Hospital Literature Review, 2006. URL `http://www.nelhin.on.ca/uploadedFiles/Public_Community/Report_and_-Publications/ALC/Literature_Review_Hospital_Dis.pdf`. Available online.

M. Nielsen, B. Langner, C. Zema, T. Hacker, and P. Grundy. Benefits of Implementing the Primary Care Patient-Centered Medical Home: A Review of Cost & Quality Results. Technical report, Patient-Centered Primary Care Collaborative, Sep 2012.

P. A. Nutting, W. L. Miller, B. F. Crabtree, C. R. Jaen, E. E. Stewart, and K. C. Stange. Initial lessons from the first national demonstration project on practice transformation to a patient-centered medical home. *The Annals of Family Medicine*, 7(3):25460, 2009.

P.A. Nutting, M.A. Goodwin, S.A. Flocke, S. Zyzanski, and K.C. Stange. Continuity of primary care: To whom does it matter and when? *Ann Fam Med*, 1:149–55, 2003.

K. J. O'Leary, D.M. Liebovitz, and D. W. Baker. How hospitalists spend their time: insights on efficiency and safety. *Journal of Hospital Medicine*, 1(2):88–93, 2006.

Y. A. Ozcan. In *Quantitative methods in health care management: techniques and applications*, volume 4, chapter Queuing models and capacity planning, pages 345–72. John Wiley & Sons, 2005.

A. Ozen and H. Balasubramanian. The impact of case-mix on timely access to appointments in a primary care group practice. *Health Care Management Science*, 16(2):101–18, 2013.

Ozen, A. and Balasubramanian, H. and Roche, J. and Samra, P. and Ehresman, M. and Li, H. and Fairman, T. Impact of discharge timings and modeling hospital-wide patient flows using simulation, 2014. Working paper.

Ozen, A. and Marmor, Y. and Rohleder, T. and Balasubramanian, H. and Huddleston, J. and Huddleston, P. Optimization and Simulation in the Orthopedic Spine Surgery Practice at Mayo Clinic, 2014a. Submitted to MSOM Special Issue on Practice Focused Research.

Ozen, A. and Marmor, Y. and Rohleder, T. and Balasubramanian, H. and Huddleston, J. and Huddleston, P. A New Spine Surgery Categorization to Improve Operating Room Scheduling., 2014b. Working paper.

J. Patrick and M. L. Puterman. Reducing wait times through operations research: Optimizing the use of surge capacity. *Healthcare Policy*, 3(3):75–88, 2008.

PCPCC. Defining the medical home, 2013. URL `http://www.pcpcc.net/about/medical-home`.

K. Phan and S. R. Brown. Decreased continuity in a residency clinic: a consequence of open access scheduling. *Fam Med*, 41(1):46–50, 2009.

B. Potts, R. Adams, and M. Spadin. Sustaining primary care practice: A model to calculate disease burden and adjust panel size. *The Permanente Journal*, 15(1): 53–6, 2011.

E. S. Powell, R. K. Khare, A. K. Venkatesh, B. D. Van Roo, J. G. Adams, and G. Reinhardt. The relationship between inpatient discharge timing and emergency department boarding. *The Journal of Emergency Medicine*, 42(2):186–196, 2012.

N.C. Proudlove, S. Black, and A. Fletcher. OR and the challenge to improve the NHS: modelling for insight and improvement in in-patient flows. *Journal of the Operational Research Society*, 58(2):145–158, 2007.

X. Qu, R. Rardin, J.A.S. Williams, and D. Willis. Matching daily healthcare provider capacity to demand in advanced access scheduling systems. *European Journal of Operational Research*, 183(2):812–826, 2006.

R. Quinn. The earlier, the better, October 2011. URL `http://www.the-hospitalist.org/details/article/1354283/The_Earlier_-the_Better.html`.

M. Ramakrishnan, D. Sier, and P. G. Taylor. A two-time-scale model for hospital patient flow. *IMA Journal of Management Mathematics*, 16(3):197–215, 2005.

Reid, R. J. and Coleman, K. and Johnson, E. A. and Fishman, P.A. and Hsu, C. and Soman, M.P. and Trescott, C. E. and Erikson, M. and Larson, E.B. The Group Health Medical Home At Year Two: Cost Savings, Higher Patient Satisfaction, And Less Burnout For Providers. *Health Affairs*, 29(5):835–843, 2010.

Reid, R. J. and Fishman, P.A. and Yu, O. and Ross, T. R. and Tufano, J. T. and Soman, M.P. and Larson, E.B. Patient-Centered Medical Home Demonstration: A Prospective, Quasi-Experimental, Before and After Evaluation. *The American Journal of Managed Care*, 15(9):71–87, 2009.

R. Resar, K. Nolan, D. Kaczynski, and K. Jensen. Using real-time demand capacity management to improve hospitalwide patient flow. *The Joint Commission Journal on Quality and Patient Safety*, 37(5):217–27, May 2011.

D.R. Rittenhouse, S.M. Shortell, and E.S. Fisher. Primary care and accountable care-two essential elements of delivery-system reform. *N Engl J Med*, 362(4):2301–3, 2009.

L. Robinson and R. Chen. A comparison of traditional and open access policies for appointment scheduling. *Manufacturing and Services Operations Management*, 12 (2):330–347, 2010.

T.R. Rohleder, D. Sabapathy, and R. Schorn. Surgical suites' operations management. *Clin Invest Med*, 28(6):353–5, 2005.

N. P. Roos, K. C. Carriere, and D. Friesen. Factors influencing the frequency of visits by hypertensive patients to primary care physicians in winnipeg. *Canadian Medical Association Journal*, 159(7):777–83, 1998.

J. W. Saultz. Defining and measuring interpersonal continuity of care. *The Annals of Family Medicine*, 1(3):134–43, 2003.

S. Savin. In Randoph W. Hall, editor, *Patient Flow: Reducing Delay in Healthcare Delivery*, chapter 5: Managing Patient Appointments In Primary Care, pages 123–50. Springer, 2006.

C. Schoen, M. M. Doty, R. H. R, and S. R. Collins. Affordable care act reforms could reduce the number of underinsured US adults by 70 percent. *Health Affairs*, 30 (9):1762–71, 2011.

Shi, P. *Stochastic modeling and decision making in two healthcare applications: Inpatient flow management and influenza pandemics*. PhD thesis, Georgia Institute of Technology, December 2013.

Shi, P. and Chou, M. C. and Dai,J.G. and Ding, D. and Sim, J. Hospital Inpatient Operations: Mathematical Models and Managerial Insights, Sep 2012. URL `http://www2.isye.gatech.edu/people/faculty/dai/publications/shi-etal-2012.pdf`. Working paper.

H. C. Sox, M.C. Higgins, and D.K. Owens. *Medical decision making*. John Wiley & Sons, 2013.

W. E. Spangler, D. P. Strum, L.G. Vargas, and J. H. May. Estimating procedure times for surgeries by determining location parameters for the lognormal model. *Health care management science*, 7(2):97–104, 2004.

B. Starfield, J. Weiner, L. Mumford, and D. Steinwachs. Ambulatory care groups: a categorization of diagnoses for research and management. *Health Serv Res*, 26, 1991.

B. Starfield, L. Shi, and J. Macinko. Contribution of primary care to health systems and health. *Milbank Quarterly*, 83(3):457–502, 2005.

A. Testi, E. Tanfani, and G. Torre. A three-phase approach for roperating theatre schedules. *Health Care Management Sci.*, 10(2):163–72, 2007.

M. D. Tipping, V. E. Forth, K. J. O'Leary, D. M. Malkenson, D. B. Magill, K. Englert, and M. V. Williams. Where did the day go? a time motion study of hospitalists. *Journal of Hospital Medicine*, 5(6):323–8, 2010.

A.L. Tucker, A.C. Edmondson, and S. Spear. When problem solving prevents organizational learning. *Journal of Organizational Change Management*, 15(2):122–37, 2001.

J. M. Van Oostrum, M. Van Houdenhoven, J. L. Hurink, E.W. Hans, G. Wullink, and G. Kazemier. A master surgical scheduling approach for cyclic scheduling in operating room departments. *OR Spectrum*, 30(2):355–74, 2008.

C. van Walraven and A.L. Weinberg. Quality assessment of a discharge summary system. *CMAJ*, 152:1437–42, 1995.

J.M.H. Vissers, I.J.B.F. Adan, and J.A. Bekkers. Patient mix optimization in tactical cardiothoracic surgery planning: a case study. *Journal of Managament Mathematics*, 16:281–304, 2005.

R. Wachter and L. Goldman. The emerging role of "hospitalists" in the American health care system. *N Engl J Med*, 335(7):514–7, 1996.

R.M. Wachter and L. Goldman. The hospitalist movement 5 years later. *JAMA*, 287 (4):487–494, 2002.

W. Wang and D. Gupta. Adaptive appointment systems with patient preferences. *MSOM*, 37(1):111–126, 2011.

M. Weinberger, Eugene Z. O., , and William G. H. Does increased access to primary care reduce hospital readmissions? *New England Journal of Medicine*, 334(22): 1441–7, 1996.

E. N. Weiss. Models for determining estimated start times and case orderings in hospital operating rooms. *IIE Transactions*, 22(2):143–50, 1990.

Larry Wellikson. You can see 40 a day. *The Hospitalist*, November 2010. ISSN 1553-085X.

Dave West. Thousands of hospital ward transfers are unnecessary. *The Journal of Emergency Medicine*, October 2010a. URL `http://www.nursingtimes.net/specialist-news/infection-control-news/troubled-hospital-makes-patient-bed-transfers-a-priority`. Available online.

Dave West. Hospital bed transfers put thousands of patients at risk of infection, October 2010b. URL `http://www.nursingtimes.net/specialist-news/infection-control-news/hospital-bed-transfers-put-thousands-of-patients-at-risk-of-infection`. Available online.

M. Williams. 2: Hospitals and clinical facilities, processes and design for patient flow. In Randoph W. Hall, editor, *Patient Flow: Reducing Delay in Healthcare Delivery*, pages 45–79. Springer, 2006.

G. Yom-Tov and A. Mandelbaum. The Erlang-R queue: Time-varying QED queues with reentrant customers in support of healthcare staffing. *Extended Abstract for the MSOM 2010 Conference*, 2010.

S. Zeltyn, B. Carmeli, O. Greenshpan, Y. Mesika, S. Wasserkrug, P. Vortman, Y.N. Marmor, A. Mandelbaum, A. Shtub, T. Lauterman, D. Schwartz, K. Moskovitch, S. Tzafrir, and F. Basis. Simulation-based models of emergency departments: Operational, tactical and strategic staffing. *Submitted to ACM Transactions on Modeling and Computer Simulation*, 2009.