Operating on Quality, Access, and Cost: Managing Better Health Systems

Aaron Hugh Ratcliffe

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Kenan-Flagler Business School (Operations).

Chapel Hill
2013

Approved by:
Dr. Ann Marucheck
Dr. Wendell Gilland
Dr. Vinayak Deshpande
Dr. Kristen Hassmiller
Dr. Vidyadhar Kulkarni
Dr. Serhan Ziya
Current spending in the US health system has reached 17.9% of GDP and the need for improved decision-making to help lower costs and improve quality is widely recognized. This dissertation, “Operating on Quality, Access, and Cost: Managing Better Health Systems” takes a hierarchical approach to effective healthcare decision-making by examining three broad areas for healthcare improvement: health systems design, health systems maintenance, and clinical operations. We examine how effective operational strategies for improving health service delivery take into account interrelationships between quality, access, and cost of care. At the design level, we employ competitive queueing models to study the impact of inter-provider competition on quality, wait-time and social welfare. At the maintenance level, we use queueing network analysis to study the relationship between screening guidelines and capacity planning for colorectal cancer. At the operations level, we employ stochastic modeling to analyze appointment allocation policies to improve outpatient clinics’ responsiveness to patients’ needs.
ACKNOWLEDGMENTS

The completion of this dissertation would not have been possible without the assistance and support of a wide community of mentors, colleagues, family, and friends.

I would like to express my deepest gratitude and appreciation to my advisors, Dr. Wendell Gilland and Dr. Ann Marucheck, for their patience, guidance, and support in completing this research and the requirements of PhD program. You have shown great generosity by giving countless hours of your time, being available to answer any question, big or small, and helping me learn the ropes as a member of the academic community. Most importantly, you have served as trusted mentors. I am blessed to be able to model my career after you for many years to come.

I would like to thank the other committee members, Dr. Vinayak Deshpande, Dr. Kristen Hassmiller, Dr. Vidyadhar Kulkarni, Dr. Serhan Ziya for their time and effort in supporting the completion of this research and for their helpful questions and constructive feedback. I extend a special thanks to Dr. Serhan Ziya for his guidance during the completion of my M.S. program in the Statistics and Operations Research Department, his assistance in transitioning to this PhD program, and his collegial support throughout this program. I would like to thank Dr. Curtis McLaughlin for serving as an honorary member of the committee and for sharing his wisdom and experience over the past 3 years.

I would like to thank the Kenan-Flagler Business School community as I have been blessed by the many people and resources available to me during my PhD Program. I would like to thank all of the faculty in the Operations area for their classroom instruction, their feedback on academic presentations, and their general advice on becoming an effective academic professional. I would like to thank my fellow PhD Students: Gocke Esenduran, Yen-Ting Lin, Vidya Mani, Karthik Natarajan, Adem Orsdemir, Olga Perdikaki, Gang Wang, and Zhe Wang. Your friendship is
some of the greatest rewards I take with me upon completion of the degree. I would like to thank Sharon Parks, Holly Guthrie, and Erin Rimmer for their assistance throughout the program.

Above all, I give thanks to my family. I give thanks to my parents, Tom and Kathy Ratcliffe, for unconditional love and for always believing I could achieve whatever I set my mind to. I give thanks to my brother Luis, for the joy, love, laughter, and stories that have carried me through the past 5 years. To my family, I love you like a world. I celebrate this accomplishment with you and because of you.
# Table of Contents

1 Introduction ................................................................. 1

2 Competition Between Health Providers: Impact on Quality, Access, Cost, and Social Welfare .................................................. 11
   2.1 Introduction ........................................................................ 11
   2.2 Literature Review ................................................................. 16
   2.3 Model Formulation ................................................................. 21
      2.3.1 Base Model Analysis: Duplication of Quality Effort ................. 26
      2.3.2 Special Case, $\delta = 0$ ....................................................... 33
      2.3.3 Sensitivity Analysis .......................................................... 34
   2.4 Fixed Price Competition: Quality Effort Reduction .................. 38
      2.4.1 Utilization-Dependent Utility ............................................. 38
      2.4.2 Technical vs. Functional Quality ......................................... 41
   2.5 Fixed-Price Competition: Quality Effort Escalation .................. 45
      2.5.1 Quality cost per arrival ...................................................... 45
      2.5.2 Capacity-Dependent Cost of Quality .................................... 47
   2.6 Single Decision Variable Model ........................................... 50
      2.6.1 Special Case: $\delta = 0$ ....................................................... 56
      2.6.2 Interpretation and Sensitivity Analysis ................................. 58
   2.7 Summary ............................................................................. 61
   2.8 Conclusions & Future Work ................................................ 62

3 Capacity Planning for Cancer Prevention .................................... 66
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Introduction</td>
<td>66</td>
</tr>
<tr>
<td>3.2</td>
<td>Literature Review</td>
<td>73</td>
</tr>
<tr>
<td>3.3</td>
<td>Model Description</td>
<td>77</td>
</tr>
<tr>
<td>3.3.1</td>
<td>Preliminary Analysis</td>
<td>84</td>
</tr>
<tr>
<td>3.4</td>
<td>Model Analysis: Capacity Decision</td>
<td>87</td>
</tr>
<tr>
<td>3.4.1</td>
<td>Cost-Minimizing Capacity</td>
<td>87</td>
</tr>
<tr>
<td>3.4.2</td>
<td>Sensitivity Analysis</td>
<td>90</td>
</tr>
<tr>
<td>3.5</td>
<td>Model Analysis: Choosing Average Time Between Screens for Fixed Capacity</td>
<td>94</td>
</tr>
<tr>
<td>3.5.1</td>
<td>Special Case, $\alpha = 0; \theta = 0$</td>
<td>94</td>
</tr>
<tr>
<td>3.6</td>
<td>Model Analysis: Joint Optimization for Capacity and Time Between Screens</td>
<td>96</td>
</tr>
<tr>
<td>3.6.1</td>
<td>Special Case, $\alpha = 0; \theta = 0$</td>
<td>97</td>
</tr>
<tr>
<td>3.7</td>
<td>Numerical Study</td>
<td>99</td>
</tr>
<tr>
<td>3.7.1</td>
<td>Estimating Parameters</td>
<td>100</td>
</tr>
<tr>
<td>3.7.2</td>
<td>Capacity Decision, Fixed Time at Home</td>
<td>104</td>
</tr>
<tr>
<td>3.7.3</td>
<td>Average Time at Home Decision, Fixed Capacity</td>
<td>109</td>
</tr>
<tr>
<td>3.7.4</td>
<td>Joint Capacity and Average Time at Home Decisions</td>
<td>112</td>
</tr>
<tr>
<td>3.8</td>
<td>Conclusions and Future Work</td>
<td>118</td>
</tr>
<tr>
<td>4.1</td>
<td>Revenue Management for Outpatient Appointments</td>
<td>124</td>
</tr>
<tr>
<td>4.2</td>
<td>Literature Review</td>
<td>128</td>
</tr>
<tr>
<td>4.3</td>
<td>Model</td>
<td>133</td>
</tr>
<tr>
<td>4.3.1</td>
<td>Setting and Notation</td>
<td>133</td>
</tr>
<tr>
<td>4.3.2</td>
<td>Model Analysis</td>
<td>139</td>
</tr>
<tr>
<td>4.3.3</td>
<td>Structural Properties</td>
<td>142</td>
</tr>
<tr>
<td>4.3.4</td>
<td>Bounds on Optimal Solution</td>
<td>145</td>
</tr>
<tr>
<td>4.3.5</td>
<td>Sensitivity Analysis</td>
<td>147</td>
</tr>
<tr>
<td>4.4</td>
<td>Numerical Study</td>
<td>150</td>
</tr>
<tr>
<td>4.4.1</td>
<td>Experimental Design</td>
<td>151</td>
</tr>
</tbody>
</table>
List of Figures

1.1 Iron Triangle: Tradeoffs of Quality, Access, and Cost of Care ................. 4
1.2 Multiple Stakeholders influence health system design and delivery ............. 7

2.1 Competition Can Increase Social Welfare if Capacity Is Fixed .................. 59

3.1 Cost of Colonoscopy Varies Across the US (Rosenthal 2013) .................. 70
3.2 Model Setting ....................................................................................... 79
3.3 Optimal Capacity Increases in Mean Time Until Onset of Cancer ............... 107
3.4 Optimal Cost Decreases in Mean Time Until Onset of Cancer .................. 108
3.5 Optimal Cost is Not Monotone in Mean Time Until Other Detection .......... 109
3.6 Optimal Cost Decreases in Mean Time Until Next Screening Request .......... 110
3.7 Cost Function Is Convex in Mean Time at Home in Feasible Region .......... 111

4.1 Attendance Rate by Appointment Delay ............................................. 125
4.2 Model Setting ....................................................................................... 135
4.3 Unimodal and Multimodal Profit Functions ......................................... 145
4.4 Optimal Policy and Profit Sensitivity Analysis .................................... 149
4.5 Sensitivity of Optimal Solution to $p/h$ when $P(D_1 \geq k) > \alpha_2$ .......... 160
List of Tables

1.1 International Comparison of Health Expenditures and Outcomes ........................................ 2

2.1 Comparative Statics: Optimal Quality, Wait Time, Profit, and Social Welfare .......................... 34
2.2 Comparative Statics: Single Decision Model ........................................................................... 60
2.3 Summary of Impact of Competition on Quality, Access, Social Welfare ................................. 61

3.1 Sensitivity of Optimal Solution .............................................................................................. 90
3.2 Sensitivity Analysis: Choose time at home for fixed capacity, $\alpha = 0; \theta = 0$ ................... 96
3.3 Sensitivity Analysis: Choose time at home and capacity, $\alpha = 0; \theta = 0$ ....................... 98
3.4 Optimal Capacity, Wait Time, and Total Cost ....................................................................... 105
3.5 Capacity Needed to Guarantee Low Wait Times ...................................................................... 106
3.6 Optimal Time at Home, Wait Time, and Total Cost ............................................................... 111
3.7 Optimal Capacity, Time at Home, Wait Time, and Total Cost - 1 ........................................ 113
3.8 Optimal Capacity, Time at Home, Wait Time, and Total Cost - 2 ......................................... 116
3.9 Optimal Capacity, Time at Home, Wait Time, and Total Cost - 3 ......................................... 118

4.1 Summary Statistics of Percent Difference from Optimal by policy .......................................... 156
4.2 Summary Statistics of Percent Increase over FCFS by policy .................................................. 157
4.3 Summary Statistics of Policy Rankings (1 highest expected profit, 9 lowest) ......................... 158
4.4 Count of Scenarios (out of 198) .............................................................................................. 158
4.5 Class 2 Booking Limit (left) and Class 1 Protection Level (right) by Policy ............................... 161
4.6 Alternative Performance Measures by Policy .......................................................................... 163
4.7 Avg. Percent Difference from Optimal by Class 2 Attendance Rate ........................................ 164
4.8 Avg. Percent Difference from Optimal by Mean Class 1 Demand ............................................ 165
4.9 Avg. Percent Difference from Optimal by Mean Class 2 Demand ............................................ 165
4.10 Avg. Percent Difference from Optimal by $p/h, (p = 100)$ ...................................................... 166
Chapter 1

Introduction

Symptoms show that the US Health System is a body in need of repair. Along three key dimensions of system performance the system functions poorly: costs are high and rising, quality is sub-par and inconsistent, and access to care is limited and highly variable. US health expenditures are higher than any other country on a per-capita basis, reaching $8,361 per-capita in 2010, and exceed any other developed country in terms of percent of GDP, reaching 17.9% in 2010. (WHO 2012). While costs of care are high in the US, many standard outcome measures of the quality of care, such as life expectancy and mortality, have fallen behind other countries - see Table 1.1. Nearly 15 years after the well-publicized Institute of Medicine report, “To Err Is Human: Building a Safer Health System” claimed that 44,000-98,000 Americans die annually due to medical errors (Linda T. Kohn and Molla S. Donaldson 2000), the effectiveness of efforts to improve patient safety remains unclear. The Centers for Disease Control and Prevention more recently estimated that 100,000 Americans die from healthcare associated infections (CDC 2002). The Office of the Inspector General estimates that 180,000 Medicare patients experience an adverse event which contributes to their death (Levinson 2010). Landrigan et al. (2010) find that 18.1% of hospital admissions experience at least one medically induced harm with 63% of all harms identified as preventable. In terms of access, patients experience long delays for outpatient appointments and crowding in the emergency room. A 2009 survey by Merritt and Hawkins shows that the average time until a routine physical with a family practice provider is 20.3 days nationwide, ranging from 8 to 63 days across 15 selected metropolitan areas. Bour-
Table 1.1: International Comparison of Health Expenditures and Outcomes

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>USA</td>
<td>17.85</td>
<td>79</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>Netherlands</td>
<td>11.96</td>
<td>81</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>France</td>
<td>11.63</td>
<td>81</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Canada</td>
<td>11.18</td>
<td>81</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>Denmark</td>
<td>11.15</td>
<td>79</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Germany</td>
<td>11.06</td>
<td>80</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Switzerland</td>
<td>10.86</td>
<td>82</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Austria</td>
<td>10.64</td>
<td>80</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Belgium</td>
<td>10.60</td>
<td>80</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Portugal</td>
<td>10.36</td>
<td>79</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Italy</td>
<td>9.50</td>
<td>82</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Sweden</td>
<td>9.36</td>
<td>81</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>9.32</td>
<td>80</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Japan</td>
<td>9.27</td>
<td>83</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Norway</td>
<td>9.07</td>
<td>81</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>


goois et al. (2008) estimate that 1.79% of all emergency department visits left without being seen from 2000-2006. Hsia et al. (2011) find that this percentage ranged from 0% to 20.3% with a median percentage of 2.6% across 262 hospitals in California in 2007. These persistent challenges, marked by medical errors, long delays for health service, and healthcare resource shortages, are prevalent throughout other developed countries as well, but the high comparative costs and low comparative outcomes in the US raise the question of whether the US is spending more to get less and why (McLaughlin and McLaughlin 2008).

Numerous explanations have been offered for these symptoms of poor system performance and for most, no single explanation is evident or sufficient. The Institute of Medicine 2001 report “Crossing the Quality Chasm: A New Health System for the 21st Century” cites several reasons for the gap between actual and potential performance of the health system: growing complexity of science and technology, increase in chronic conditions, poorly organized delivery system, and constraints on information technology adoption. Newhouse (2002) expands on these exploring 5 possible causes of poor performance: consumer ignorance, rate of technological change,
administered prices, difficulty of measuring provider performance, and the role of the public sector. Porter and Teisberg (2006) argue that competition on the wrong levels has failed the US health system as evidenced by rising costs, high variation in quality across geographic areas, and slow technological innovation. The authors contend that the current zero-sum competition, where system participants work to gain bargaining power and shift costs to one another, leads to competition which rewards breadth of services as opposed to treatment of specific medical conditions, discrete interventions as opposed to the full cycle of care, and local bias as opposed to performance accountability. Bohmer (2009) proposes three main problems: 1) we don’t know what to do, i.e. shortage of evidenced-based practice of medicine; 2) we don’t do what we know, i.e. proper care is under or over-supplied; 3) when we do what we know, we don’t do it right, i.e. medical errors and patient safety remain critical concerns. Green (2012) cites several causes of the medical dilemma including: misaligned payment systems, high levels of uninsured or under-insured patients, overloaded physicians, profit seeking insurers, inefficient and highly fragmented health systems, and lack of communication and coordination.

In order to evaluate potential interventions and decide on the best treatments for improving the system, it is important to better understand three key dimensions of health system performance, quality, access, and cost, and the interrelationships between them. In the healthcare services context, we can define quality as the degree to which health services increase the likelihood of the desired outcome. In health services it is important to distinguish functional quality, experiential elements of service such as amenities and reputation, from technical quality, i.e. clinical outcomes, especially due to information asymmetry and principal-agent relationships between patients and providers, (Gronroos 1993). Healthcare service consumption can be seen as a credence purchase in that patients often use functional quality as a signal for technical quality. Access can be defined as the degree to which services can be obtained. Access may refer to insurance coverage in the population, disparities in care due to socio-economic status, or the distance patients must travel to obtain a particular service. Throughout this dissertation, we will measure access primarily by how long patients must wait for a health service. In terms of cost, different stakeholders have different perspectives on the cost as providers are concerned with internal and external costs of supply, insurers are concerned with reimbursement prices,
The “Iron Triangle” of quality, access, and cost refers to a common paradigm for explaining how improving performance along any one of these three key dimensions may compromise performance on one or both of the other dimensions (Kissick 1994). For example, increasing access to care in a system may require investing in additional resources (higher costs) or faster service speed which could lead to lower quality. Such tradeoffs are not always required, for example changes to payment structures, technological innovation, or lean improvement efforts to reduce waste could provide higher value (higher quality at lower cost) (DOJ and FTC 2004). However, at some point tradeoffs between quality, access, and cost become fundamental to decisions made by stakeholders at the design, maintenance, and operational levels of health systems.

Decisions regarding how to balance quality, access, and cost are made across multiple stakeholders at multiple levels, each with its own priorities. Patients are the primary stakeholders in that it is their individual health which is in question, but patients differ in preferences, values, and needs. For example, some consumers may prefer “nothing but the best” while other consumers may be willing to exchange lower quality service for lower prices or wait longer for...
higher quality service (DOJ and FTC 2004). As stakeholders, patients balance quality, cost, and access through questions such as: Should I seek care with a lower ranked specialist who is more convenient or affordable? Should I follow guidelines for preventive cancer screening? Should I schedule an appointment well in advance or wait until I know better my availability? Patients often have little individual influence within the system, leaving much of the decision-making power to larger stakeholders.

Providers are also stakeholders and face issues such as: How much should we invest in quality and process improvement efforts? How much capacity do we need to meet demand? What mix of services should we offer? Should we merge with another competing provider? How do we schedule patient encounters and resources? State and federal governments are also a key stakeholders and faces issues such as: Should hospitals be allowed to operate as monopolists or is provider competition better for society? Which services, e.g. preventive cancer screenings, should be covered by Medicare and Medicaid? Where should funds be invested in disease prevention, and healthy behaviors? Payers offer insurance and financing options for patients and employers who rely on healthy, productive employees. Within the supply chain, stakeholders such as pharmaceutical companies and medical equipment suppliers influence the cost of care, the rate of new product development, and the diffusion of technological innovation. With so many stakeholders, movements such as patient-centered care attempt to unite stakeholders and actively involve, educate, and engage patients in decisions which respect their preferences.

Recent major health policy reform in the US, via the Patient Protection and Affordable Care Act (ACA) of 2010 has drawn even greater public attention to the concerns surrounding our health care delivery system, and the reform promises to bring with it significant changes to the system (Assistant Secretary of Public Affairs 2010). One significant, and commonly cited change is the provision for increased health insurance coverage which requires all citizens to carry health insurance or pay a federal tax penalty. The increase in health insurance coverage is expected by many to lead to an increase in demand for health services as previously uninsured or under-insured individuals will now be required to have coverage. Combined with the increased demand for health services due to aging of the baby boomer generation, health providers must
plan accordingly for increased patient volume. The impact of an increased demand is a recurring question across different types of health services and levels of analysis.

A second significant change is the set of provisions within the ACA which emphasize preventive services. In particular, a new health insurance plan or insurance policy beginning on or after September 23, 2010, must cover 15 preventive services for adults, 22 for women, and 26 for children without the patient having to pay a co-payment or co-insurance or meet a deductible when the services are delivered by a network provider (U.S. Department of Health & Human Services 2013). The increase in demand for preventive health services could therefore be even greater. Colonoscopy screening for patients over 50 is an example of such a preventive service which has seen a dramatic increase in volume in the past 10 years (Rosenthal 2013).

A third significant development connected to the ACA is the emphasis on accountable care organizations (ACO) and the idea of a patient-centered medical home. In an ACO, a provider’s reimbursement is linked to quality of care metrics for managing the overall health for a given population (Centers for Medicare & Medicaid Services 2013). The ACA authorizes the Center for Medicare and Medicaid Services (CMS) to create shared savings programs to contract with providers. The benefits of ACOs are that they alter provider incentives so that there is a greater emphasis on value for the patient (higher quality at lower costs), and they support increased efficiency and coordination across different levels of care (e.g. primary, specialty, etc.). ACOs are not without challenges. There are a lack of clear guidelines for implementation and issues concerning provider consolidation and competition. While the federal government indirectly encourages providers to form ACOs via consolidation under the ACA, the Federal Trade Commission (FTC) still encourages market competition and providers who form ACOs run the risk of anti-trust infringement due to consolidation. Better understanding of the impact of provider consolidation on quality, cost, and access to service is warranted in order to effectively balance these two opposing pressures relating to implementation of ACOs.

A fourth important change related to the ACA is information transparency and the creation of health information exchanges in order to provide patients with better information to make health service consumption decisions. In order for patients to accurately balance the tradeoffs
between quality, cost and access described in the Iron Triangle, the patients must have access to accurate information about each of the dimensions across competing providers. It is unclear how patients will actually utilize public reporting of provider quality, cost, and service delivery metrics, but the general agreement is that improved consumer access to such information places greater pressure on providers to deliver higher quality service, with better access at lower cost.

In this dissertation we use operations management theory and analysis to investigate three critical research problems which correspond to a three-level hierarchy of health systems analysis: health system design, health system maintenance, and health system operations. At the design level, we employ competitive queueing models to further investigate the role of competitive structures in health care. By examining the role of competition, we study one important proposed explanation for poor system performance. We provide insights regarding whether policy makers should use programs such as Certificate of Need to regulate competition for health services. We also investigate how provider strategy, and the balance of quality, cost, and access, changes from a monopoly to a duopoly environment. At the maintenance level, we use queueing network analysis to study the relationship between screening guidelines and capacity planning.
for colorectal cancer. Lack of emphasis on prevention is another critical explanation for poor system. By examining a research problem related to colorectal cancer, we provide insight and contribute to research on planning supply and demand for preventive services. At the operations level, we employ stochastic modeling to analyze appointment allocation policies to improve outpatient clinics’ responsiveness to patients’ needs. These three research problems by no means represent the exhaustive resume of how operations research has been applied in healthcare, but they are a selection of problems which are united by 1) their focus on health service delivery and managing the associated operations as opposed to problems which emphasize medical decision making; and 2) their emphasis of the tradeoffs between quality, access, and cost and how to extend the boundaries of the Iron Triangle.

In Chapter 2, we take a design-level perspective to conduct a theoretical analysis of how competitive structures impact system performance, in terms of quality, access, cost and social welfare with fixed prices. We use generalized service models to analyze the setting within the health services industry. During the 1970s and 1980s, when price competition in health services delivery was fairly weak, analysts argued that provider competition was based on non-price elements of the service (Gaynor 2006). Anecdotes and scholarly claimed that this led to an escalation of technology and capacity investments termed the Medical Arms Race. Prior economic and health policy research, both theoretical and empirical, has demonstrated that competition increases quality when prices are fixed but has largely overlooked the impact on access to care, in particular wait-time dynamics (Gaynor 2006). We analyze a competitive queueing model, an increasingly common approach in general service provider equilibrium models in the literature, to compare equilibrium profit, social welfare, quality effort and capacity decisions between a monopoly and duopoly setting. We model health providers as single-server queues that maximize expected profits by optimizing a quality effort and capacity decision. Consumers select a provider to maximize utility, which is a function of provider quality effort, co-payment, and expected wait time. In our base model, we find that competition leads to a duplication of quality effort and capacity buffers and a corresponding loss of social welfare when providers choose both quality effort and capacity. However, by modeling different cost functions, wait functions, and types of service quality, we show that competition may lead to a reduction or escalation of
quality effort. When providers only choose quality effort (capacity is fixed), competition may lead to lower quality and lower wait-time, and an increase in social welfare. The models which show a reduction of quality effort are significant in that they counter previous economic theory and the Medical Arms Race argument that price-regulated competition always leads to higher quality.

In Chapter 3, at the health maintenance level, we examine screening policy and capacity allocation decisions for Colorectal Cancer (CRC) prevention. CRC is a significant health concern both in the US and worldwide. The National Cancer Institute estimates 146,970 new cases in 2009 with 106,100 of those being colon cancer (National Cancer Institute 2013b). CRC is also the second leading cause of cancer death in the United States (National Cancer Institute 2013b). We examine two research questions: 1) How should screening providers or policy makers design their screening capacity to minimize the sum of capacity and detection delay costs? 2) What are the tradeoffs between the optimal guidelines for time between screening and the optimal capacity? We use a queueing network model to capture patient return behavior for cancer screenings. A social planner selects a capacity to optimize expected total cost per unit time which is the sum of expected detection delay costs (for all patients in the population) and the cost of capacity. We calculate the optimal capacity assuming the average time until the next screening request is exogenous. In order to better understand the tradeoffs between screening guidelines and screening capacity, we analyze a model where the provider optimizes the average time until the next screening request for fixed capacity, and also investigate a joint model where the provider optimizes both the average time until the next screening request and the capacity. Given the ACA emphasizes colonoscopy as a preventive service which new insurance policies must cover at no expense to the patient, understanding capacity planning for colonoscopy is a relevant and noteworthy problem.

In Chapter 4, at the operational level, we study how clinics should allocate appointment capacity to improve outpatient appointment access and reduce the wasteful impact of no-shows. Poor appointment access is a growing trend. In 2001, 33% of patients reported an inability to obtain a timely appointment as opposed to 23% in 1997 (Strunk and Cunningham 2002).
Simultaneously, clinics experience alarmingly high no-show rates, and research supports that no-show rates are higher when patients schedule further in advance (Gallucci et al. 2005, Green and Savin 2008, Liu et al. 2010). Clinics must balance scheduling advance requests at high no-show rates and holding appointments open for potential immediate requests at low no-show rates. Open Access is a popular paradigm for clinics to “do today’s work today” by offering same-day appointments and encouraging short booking windows. As encountered by clinics in the UNC Health Care System Patient Access and Efficiency Initiative, Open Access does not eliminate the underlying challenges of managing clinic capacity since some patients request advance appointments, due to preferences, medical condition, travel arrangements, or ease of scheduling follow-up visits. We analyze how a clinic should control bookings from two sequential patient classes with different no-show rates in order to maximize expected profit. We employ stochastic comparisons to establish structural properties and develop approximations for the optimal appointment allocation policy. Our numerical study compares our approximations with the optimal policy and policies used in practice and previous literature. On average, we find that the optimal policy increases profits 17.8% over FCFS. A simple policy we develop performs 0.3% below optimal. “Pure” Open Access can achieve optimality in some situations but performs 23.0% below optimal on average. We provide methods for effectively implementing joint capacity control and overbooking for outpatient appointments.

There is great opportunity to improve decision-making within the health services system using analytical tools and operations management theory. The impetus for doing so is expanding for several reasons: practitioners are becoming increasingly aware of the benefits of analytical methods; the demand for health services is growing due to an aging population and the ACA; policy makers are trying to cut healthcare costs due to sluggish economy and large government deficits; and competitive pressures and regulations require improved quality performance (Green 2012). This dissertation offers insight and tools for improved decision making for both health service system managers and policy makers with respect to three important health service delivery research problems where quality, access, and cost must be balanced.
Chapter 2

Competition Between Health Providers: Impact on Quality, Access, Cost, and Social Welfare

2.1 Introduction

US health expenditures are higher than any other country on a per-capita basis, reaching $8,361 per-capita in 2010, and exceed any other developed country in terms of percent of GDP, reaching 17.9% in 2010. (WHO 2012). While costs of care are high in the US, many standard outcome measures of the quality of care, such as life expectancy and mortality, have fallen behind other countries (WHO 2010). These numbers raise the question of whether the US is spending more to get less and why. In the US, competition plays an important role in how providers balance priorities of quality, access, and cost and which strategies they adopt in order to try and improve performance on multiple dimensions. Features of the market for health services make it distinct from other markets and can limit the benefits of competition (Porter and Teisberg 2006). The question remains whether competition within the US health care system is a problem or a solution or both. We develop an analytical model of health provider competition to study provider operations strategy and the impact of competition on quality, cost and access to care.
The “iron triangle” of quality, access, and cost refers to a common paradigm in health policy analysis where improving performance along any of the three dimensions may compromise one or both of the other dimensions (Kissick 1994). Some strategies and innovations break the boundaries of these tradeoffs by improving performance in multiple dimensions, e.g. lean improvement aims at providing higher quality at lower cost by eliminating waste in the provider’s delivery system; but ultimately, these tradeoffs are necessary to the design of the health system. Additionally, these tradeoffs must be weighed across multiple levels and stakeholders; e.g. some consumers may prefer “nothing but the best” while other consumers may be willing to exchange lower quality service for lower prices or wait longer for higher quality service (DOJ and FTC 2004). To understand how competition impacts system performance as measured by these dimensions, it is fundamentally important to incorporate such tradeoffs into our model of the health service delivery market.

“Healthy” competition for health services is limited by unique features of the market including: a highly regulated marketplace; distorted incentives due to third-party payment (health insurance); information problems such as asymmetry, reliability, and lack of transparency; societal attitudes regarding medical care; and agency relationships (DOJ and FTC 2004). Additionally health service consumption can be defined as a credence purchase in that consumers often use perceptions and functional quality, experiential elements of service such as amenities and reputation, as signals for technical quality, i.e. clinical outcomes (Gronroos 1993). Porter and Teisberg (2006) contend that in a normal market competition drives relentless improvements in quality and cost, whereas competition on the wrong levels has failed the US health care system as evidenced by rising costs, high variation in quality across geographic areas, and slow technological innovation. By competition on the wrong levels, the authors refer to zero-sum
competition where system participants work to gain bargaining power and shift costs to one another. This leads to competition which rewards breadth of services as opposed to treatment of specific medical conditions, discrete interventions as opposed to full cycle of care, and local bias as opposed to performance accountability standards set by competing providers in more distant geographic regions. The authors argue, however, that while competition on the wrong levels is the core problem, competition on services which create value for patients is the core solution. As a result of the barriers to price competition in the health services market, health economists and health policy researchers have argued for the existence of a Medical Arms Race (MAR) which is characterized by an escalation of high-cost technology and capacity investments due to provider competition on non-price elements of service such as advertising, amenities, and reputation which may not lead directly to better medical outcomes (Joskow 1980, Robinson and Luft 1985, 1987, Dranove et al. 1992).

The Medical Arms Race (MAR) argument was first proposed in the 1970s and 1980s, when prices were largely regulated by the government. Since that time several significant changes have occurred in the health care landscape. The first significant change is that the onset of managed care in the mid-1990’s gave more pricing power to payers to contract with health service providers. As a result, consolidation amongst US health providers has surged since the mid-1990’s due to greater financial pressure to cut costs and increase quality with the onset of managed care. In theory, as payers developed more negotiating power, they could act as agents for the consumers and leverage greater price competition between providers. Many urban areas are now dominated by 2-3 health systems. Levin Associates estimates 900 hospital mergers and acquisitions from 1991 to 2000, and though the number dropped to 589 from 2001 to 2010, the estimated value of these mergers still exceeds $80 billion (Gaynor 2006, Irving
Levin Associates, Inc. 2011). The debate continues regarding what happens to price, quality, and access to care due to increased consolidation. Will consolidation lead to increased prices and reduced access to care as providers exert monopoly power or will it create efficiency gains and reduce unnecessary duplication of services leading to lower costs and greater affordability? By controlling the “medical arms race” through consolidation, will the overall quality of health care improve and at what cost?

A second important change is that interventions have been enacted to limit competition in healthcare markets, and debate continues regarding their usefulness. At first on the federal and then on the state level, Certificate of Need (C.O.N.) and other similar programs were developed to coordinate planning of new capacity and technology adoption, and to control high-cost expenditures deemed “medically unnecessary.” While C.O.N. was repealed on the federal level in 1987, due to its ineffectiveness in controlling costs, C.O.N. programs administered at the state level may require health providers to obtain government authorization before offering high-cost services (e.g. bone marrow transplant), investing in high-cost technologies (e.g. proton beam therapy, gamma knife or Magnetic Resonance Imaging), or changing bed capacity.

A third change relates to the evolution of consumer access to information. The information technology boom now makes it possible for consumers to access more information before selecting a healthcare provider. Using websites such as The Department of Health and Human Service’s Compare Hospital tool, consumers may compare hospitals with respect to a variety of different performance measures of process, outcomes, patient satisfaction, and Medicare reimbursement (HHS.gov). Additionally, many health providers now provide real-time wait time information, in particular for Emergency Departments, via billboards, text messages, and internet sites in an attempt to emphasize wait-time access as a strategic priority and lure customers away from
competitors (Fleisher 2011). The Affordable Care Act of 2010 (ACA) emphasizes information transparency as an important part of health policy reform with the creation of health information exchanges in order to provide patients with better information to make health service consumption decisions. Questions remain to be answered as to how health care consumers truly use this information and how it affects their choices of providers.

In light of the changing landscape, some researchers have argued for the existence of a “new medical arms race.” Devers et al. (2003) observe a shift in provider’s strategic emphasis between 1996-1997 and 2000-2001. They find that in the mid-1990s hospitals primarily competed on price by providing services attractive to managed care plans, but by 2000-2001, non-price competition was becoming increasingly important and hospitals were reviving strategies which provide services attractive to individual physicians and the patients they serve. They point out some important differences between the new medical arms race and that of the 1970s and 1980s: the hospital market is more concentrated and price competition remains relatively important. Numerous national and local news reports continue to provide anecdotal evidence of hospital rivalry and non-price competition (Whelan and Langreth 2009, Warner 2011, Emanuel and Pearson 2012, Wall 2012, Locke and Frank 2012).

In this work we analyze a competitive queuing model to examine how competition between health providers impacts quality, wait-time access, and cost under fixed prices. We compare consumer surplus, industry profit, and social welfare between a monopoly and duopoly setting. Previous economic theory consistently supports that quality will increase under fixed-price competition and lead to cost escalation. We seek to understand if the escalation is mitigated or exacerbated in a model which incorporates consumer equilibrium waiting behavior. We also perform comparative statics to study how current trends in the US market for health services,
such as an increasingly informed consumer population and growing demand, affect the role of competition. The remainder of this study is organized as follows. In Section 2, we provide a review of the relevant literature. In Section 3, we describe the model and provide results for the base assumptions. In Section 4, we analyze variations of the model where fixed-price competition leads to lower quality effort and lower wait time, a counter-argument to the Medical Arms Race. In Section 5, we analyze variations of the model where fixed-price competition leads to higher quality effort and higher wait time, supporting the Medical Arms Race argument. In Section 6, we consider a model where quality and service rate are linked by a single decision and compare the results with the base model. In Section 7, we summarize the results of the various models. In Section 8, we offer conclusions and insights from our work and discuss opportunities for future research.

2.2 Literature Review

Our work is related to theoretical and empirical research from economics and health policy as well as competitive queueing models from operations management literature.

Gaynor (2006) reviews the theoretical and empirical literature on the impact of competition on quality. Under fixed prices, he finds consistent support for higher quality and higher consumer welfare but ambiguous effects on social welfare. White (1972) and Douglas and Miller (1974) both develop models motivated by the regulated airline industry to show that when prices are regulated, firms compete away profits by offering higher quality. Held and Pauly (1983) present a simple model, motivated by study of the Medicare program for persons with end stage renal (kidney) failure, to show a “competitive amenity bias” - i.e. the level of amenity can be affected by the level of reimbursement and the level of competition. Pope (1989) examines hospital
non-price competition in Medicare’s Prospective Payment System and finds that hospitals raise their quality to compete for Medicare patients. The higher quality expenditures lead to increased costs, but simultaneously reduce slack. Allen and Gertler (1991) analyze a theoretical model where consumers are heterogeneous and a firm can endogenously discriminate based on quality under two pricing scenarios: 1) the consumer is the payer, 2) the payer is not the consumer. The authors find that fixed-price regulation leads to distributional welfare loss and show that it cannot induce providers to supply all consumer types with first-best quality under either pricing scenario. Calem and Rizzo (1995) develop a variant of the Hotelling location model where reimbursement prices are exogenously determined by third-party payers to examine hospitals which compete with respect to specialty mix and quality of service. They find competing hospitals differentiate specialties too much or too little compared to the socially optimal service mix chosen by the monopolist, thereby suggesting that hospital mergers lead to efficiency gains since the consolidated hospitals do not have to maintain excess capacity. They also find that higher reimbursement levels lead to intensified quality competition, and therefore higher costs, but also to increased service differentiation as hospitals try to relax quality rivalry. Brekke et al. (2010) use a differential-game approach to show this effect can be exaggerated in static models. Our work contributes to this stream of economic theory by considering quality and capacity competition in a fixed-price market where consumers are sensitive to wait-time. We study whether the wait-time dynamics mitigate or exacerbate the impacts of competition on quality and social welfare. Many of the previous papers from this stream consider the impact of competition on quality, but do not explicitly model access which we capture through queueing elements of our model.

From the empirical literature, numerous studies consider Medicare patients, to study a “fixed-
price” population, and show that competition, measured by the Herfindahl-Hirschman Index (HHI), improves quality, measured by decreased mortality (Gaynor 2006). Kessler and McClellan (2000) find that before 1991, hospital competition led to higher costs and, in some cases, lower rates of adverse outcomes for elderly Medicare patients with heart disease; whereas after 1991, hospital competition led both to substantially lower costs and significantly lower rates of adverse outcomes. Shen (2003) studies how the number of hospitals interacting with Medicare affects AMI mortality and finds no effect from 1985-90, but an increase in quality (decrease in mortality) from 1990-94. Tay (2003) studies the impact of demand elasticity on mortality for Medicare heart attack patients and finds an increase in quality (decrease in mortality). The authors argue that this evidence supports the claim that hospital competition improved social welfare. Kessler and Geppert (2005) find that increased competition leads to lower readmission rates for Medicare patients. Many such studies emphasize outcome-based metrics for quality and make few conclusions regarding functional elements of quality. One exception is (Held and Pauly 1983). While outcome measures of quality may be increased due to competition, it is still unclear how value is affected, i.e. whether the higher quality comes at substantially higher cost.

When prices are set by firms, the effect of competition on quality and welfare is unclear. Theoretical results suggest that competition may not necessarily lead to optimal quality effort. Kamien and Vincent (1991) and Ma and Burgess (1993) show that unregulated competition leads to suboptimal quality under exogenous product variety. Spence (1975) shows that the monopolist supplies a socially optimal level of quality only when the marginal consumer is the average consumer. Allard et al. (2009) derive conditions where physicians supply optimal effort in a repeated game, but sub-optimal effort in a static game. The empirical literature also shows conflicting results on the effect of competition on quality, even when varying measures
of quality and competition are used. Propper et al. (2004) and Propper et al. (2008) show increased competition, as measured by the number of competitors, leads to decreased quality as measured by mortality. Ho and Hamilton (2000) show that decreased competition, as measured by the number of hospital mergers, can lead to lower quality, as measured by the number of readmissions for heart attack and stroke. Thus, the effect of competition on quality in health care is an open question, particularly when prices are not fixed.

Our work is also related to equilibrium queueing models from service operations. Hassin and Haviv (2003) offer a comprehensive review of the literature on equilibrium behavior of customers and servers in queueing systems. These models largely derive from seminal work by Naor (1969) who finds that a levying toll can induce strategic customers in a queue to behave in a socially optimal manner. More recent work expands upon that model by examining the tradeoff between service rate and service quality, which has been modeled both in terms of service value and in terms of service outcomes. Kostami and Rajagopalan (2009) develop a model where a firm may speed up a process to meet more demand with less congestion, but this may result in decreased quality. They highlight the significance of this tradeoff in the healthcare sector. They also characterize and compare the equilibrium outcomes for a single provider in a single period and dynamic, multi-period setting. Anand et al. (2011) describe services where the quality-speed tradeoff is critical and demonstrate that the customer-intensity of the service is a critical driver of equilibrium price, service speed, demand, congestion in queues and service provider revenues. Pac and Veeraraghavan (2010) analyze expert service providers’ pricing and diagnosis strategies when there is information asymmetry between the expert and the consumer. The expert prices two treatments, receives asymmetric information from a diagnosis, and then refers the consumer for treatment. The authors find that congestion concerns mitigate expert cheating.
and that experts charge high prices to signal honest diagnoses. Dai et al. (2012) model test-ordering behavior for a single physician and a pool of insured patients in an outpatient setting using an equilibrium queueing model. They model service quality as diagnostic certainty which decreases in service rate (i.e. more diagnostic testing leads to diagnostic accuracy) and increases patient utility. Under the baseline model, physician over-testing always occurs due to insurance coverage and the authors also consider the effects of five different service environments. We consider similar quality-speed tradeoffs but our model takes a higher-level perspective in order to gain insights into health operations strategy; therefore, service providers represent health systems or competing clinics, either independent or from across competing health systems.

Other research from competitive queuing models examines competition between the service providers. Chen and Wan (2005) study how market size affects market structure when service providers choose price and capacity with homogeneous customers. They characterize the Nash equilibrium and show that when it is profitable for both providers to enter, there exists a continuum of equilibria where providers split the full market and adjust their respective capacity decisions accordingly. They also show that a monopoly is socially optimal and that a duopoly results in a loss of social welfare due to the dominance of economies of scale. Allon and Federgruen (2008) compare three models of competition based on the timing of price and delay decisions (delay first, price first, simultaneous) and find that choosing delay first leads to highest wait time and lowest prices. Afanasyev and Mendelson (2010) compare a generalized delay cost structure with a traditional additive delay cost structure in a model with heterogeneous customers where two competing service providers choose arrival rates and capacities. They show that when customer service valuations are additive (i.e. independent of disutility from waiting), both providers may offer differentiated services, in terms of price and delay, but customers are
always indifferent between the providers, i.e. there is no market segmentation. Under the generalized delay cost structure, where service valuations and delay costs are interdependent, they find value-based segmentation in the market when the service providers have different costs, i.e. one provider gives fast service to high-value consumers and the other gives lower price and slower service to low-value consumers. Many of the previous papers from this stream examine models with provider pricing. Our work is a contribution to previous papers on service provider competition, as it considers a fixed price model where consumers are insured, and in our model, the provider chooses its quality effort and service rate with possible inter-dependencies between the two decisions.

2.3 Model Formulation

We formulate a two-stage competitive queueing model with 2 service providers and a population (enrollment) of \( N \) potential consumers of a medical service. Here providers represent competing health systems or competing clinics, either independent or across health systems. In the first stage, Provider \( i = 1, 2 \) maximizes its long-run expected profit per unit time, \( \pi_i \), by selecting its quality effort, \( x_i \), and capacity, \( \mu_i \). Each provider incurs variable costs \( C_i(x_i, \mu_i) \) which are jointly convex in quality effort and service capacity, respectively. Each provider earns a reimbursement price \( r_i \) per consumer served. The quality effort decision represents per-period efforts made by the provider over a base quality level which may increase consumer service valuation (e.g. higher functional/experiential quality, or fewer side-effects associated with treatment) or the probability of a successful service outcome (e.g. higher technical/clinical quality), or both. However, quality effort comes with a cost. This cost can be either the financial cost associated with quality improvement projects, operating additional amenities, or the time cost, as we will
see below, associated with safety inspections or spending more face-time with the patient. The
service rate represents the aggregate capacity of the provider (clinic/facility), not necessarily
the service rate of an individual resource. Therefore, in this model, the queue does not describe
a short-term imbalance for a given resource, but an approximation for access to the service in
terms of aggregate provider utilization. The variation in arrivals and service described by the
queueing model then refers to an aggregate for the provider (e.g. patients per day, per week,
etc.). The provider can increase its capacity through means such as process improvement, hiring
additional staff, or leasing or amortizing payments on additional equipment, all of which are
aggregated under an cost of service capacity in each period.

In a given time period, consumers from the fixed enrollment require (seek) medical service
independently with probability \(s\) and generate a Poisson arrival process with arrival rate \(\Lambda = Ns\).
We assume that consumers are able to observe whether or not they need medical service and the
probability of seeking service, \(s\), is independent of any decision made by the provider, i.e. the
provider cannot induce demand from patients who are not ill (enough to seek care). Arriving
consumers observe the quality effort, the expected wait-time, and the out-of-pocket payment
at each provider a priori. For our analysis, we assume all consumers are homogenous but
specify a general model where consumer valuation for service may vary within the population.
The expected utility that consumer \(n\) receives from service at Provider \(i = 1, 2\) is denoted by
\[ U_{in} = V_{in}(x_i) - hW_i(\mu_i, \lambda_i, x_i) - p_i \]
where \(V_{in}(x_i)\) is the service valuation of consumer \(n\) at Provider \(i = 1, 2\), \(W_i(\mu_i, \lambda_i, x_i)\) is the expected throughput time at Provider \(i = 1, 2\), and \(h\)
is the disutility per unit time spent in the system. For simplicity we assume the out-of-pocket
payment paid by the consumer at Provider \(i = 1, 2\) is the same for all consumers and this price
is also exogenously given. We assume that \(V_{in}(x_i)\) is a random variable which is independent
and identically distributed across the population, stochastically increasing in the quality effort decision, $x_i$, and follows a known distribution with c.d.f. $F_{x_i}(v)$. We assume that consumers know their own service valuation as well as the service value distribution of other consumers. Given this information, consumers maximize their expected utility by deciding to join a queue (e.g. take the first available appointment) or take an outside option which has exogenous utility $b$. While the provider valuations may be heterogeneous across patients, we assume the value of the outside option is constant across all patients. In the healthcare context, the outside option might refer to emergency care for those who cannot access primary care services, destination providers such as Mayo Clinic for those who cannot access local specialty services, or home remedies for non-acute services. We are modeling healthcare settings, such as outpatient care or scheduled surgeries, where a patient's condition is not so emergent that there is no time to choose between providers or the outside option. The consumer joining decisions lead to the effective arrival rates at each provider denoted by $\lambda_i$ for $i = 1, 2$.

Using the above assumptions and notation we write the objective function for Provider $i = 1, 2$ as:

$$\max \pi_i = r_i \lambda_i - C_i(x_i, \mu_i) \quad (2.1)$$

The consumer choice model is given by:

$$U_n = \max \{V_{1n}(x_1) - hW_1(\mu_1, \lambda_1, x_1) - p_1, V_{2n}(x_2) - hW_2(\mu_2, \lambda_2, x_2) - p_2, b\} \quad (2.2)$$
Before proceeding to the model analysis, we introduce some simplifying base case assumptions and discuss limitations of our model. First we assume that costs are linear, i.e. \( C_i(x_i, \mu_i) = q_i x_i + c_i \mu_i + k \) where \( q_i \geq 0 \) denotes the linear cost of quality effort per period, \( c_i \geq 0 \) denotes the linear cost of capacity per period, and \( k \geq 0 \) denotes a fixed operating cost per period. This assumption is primarily made for convenience as either economies of scale or dependencies between the costs of quality effort and capacity may exist in reality which could mean the variable costs could be non-linear or the capacity operating cost be quality-effort dependent. We assume that both quality effort and capacity are unbounded continuous decision variables. Though it has limitations, discussed below, the linear cost function assumption is common in previous literature on service provider competition. (Chen and Wan 2005).

Second, we assume that service times are exponential so that each provider may be modeled as an M/M/1 queue. We assume that the effective mean service rate, \( \tau_i \), is linear in the provider’s chosen service rate and quality effort, i.e. \( \tau_i = \mu_i - \delta x_i \), where \( \delta \) denotes the correlation between quality effort and effective service rate. When \( \delta = 0 \), the service rate is independent of the quality effort decision. If \( \delta > 0 \), then higher quality effort may require the provider to slow its effective service rate, e.g. by spending more time serving the consumer, which may be true of amenities like noise abatement and educational programs. If \( \delta < 0 \), then increasing the quality effort further increases the provider’s service rate, e.g. the provider expands its quality effort by operating clinic management technology, such as Electronic Medical Records and computer based ordering which increase service rate and consumer service valuation. Given this assumption, we write \( W_i(\mu_i, \lambda_i, x_i) = \frac{1}{\mu_i - \delta \mu_i - \lambda_i} \). The exponential service assumption is made for mathematical tractability, however, since we are using the queueing model to represent aggregate variation in arrivals and service, the exponential assumption may over-estimate true
variation in capacity. As discussed by Green and Savin (2008), there likely is some variation in provider capacity. This variation is not only variation due to time spent with individual resources (e.g. check-in, nurse, physician etc.), but also variation in supply due to wasted appointments (no-shows, late cancellations and reschedules) and variation in number of appointment slots requested. We can generalize the results to M/G/1 queues and approximately to G/G/1 queues using the Pollaczek-Khinchine and Kingman’s formulas, respectively.

There are several limitations of the model which should be discussed before proceeding. First, we assume that the out-of-pocket payment at a given provider is the same for all consumers. In reality, consumers may pay different prices because they are on different plans (even with same insurer) or have different amounts of deductible remaining. A second limitation is that our model does not incorporate into patient preferences any disutility from switching providers which may exist when patients seek service for multiple episodes of care from the same facility. Also, our modeling of the endogeneity of demand for health service is limited. In reality, the quality of service a patient receives today may affect their future health state and demand for health services tomorrow. Consumers who receive high quality care may be more likely to return to a given provider due to a positive experience, but may not return for service for a longer period of time if the care received keeps them well longer. A third limitation is that consumers may be heterogeneous on multiple components of their utility, for instance some classes of patients may have higher sensitivity to waiting or higher expected service times. Additionally, there may exist correlations between the components of the utility function, e.g. consumers who value service more may be more sensitive to waiting. Lurking variables such as a consumer’s underlying health state might also impact multiple components of the utility function. A final limitation of the model is that the queueing service discipline does not accurately reflect appointment
scheduling systems used by many outpatient providers in practice. Appointment systems help
providers balance variation of in-clinic waiting time, reserve capacity for urgent requests, and
match supply with patient preferences. We assume a FCFS service discipline though in reality
consumers may not always prefer the first available appointment. Since we are concerned with
overall expected wait-time, we do not model the appointment schedule exactly, but acknowledge
it as an important component in many healthcare contexts. See Green and Savin (2008) for
further discussion of using FCFS queueing model for appointment delay in the outpatient setting.

Several of these limitations will be revisited by relaxing model assumptions in later sections or
extensions. While relaxing assumptions to consider various extensions provides for interesting
problems, the base case assumptions allow for a simple model to address our core research
question regarding the impact of competition.

2.3.1 Base Model Analysis: Duplication of Quality Effort

We assume that consumers are homogenous with linear quality sensitivity. This assumption
implies that the consumer service valuation at provider $i = 1, 2$ is identical for all consumers
and takes the form $V_{in} = a_i + \beta x_i \ \forall n$. The base service valuation at provider $i = 1, 2$ when
quality investment is zero is given by $a_i \geq 0$, and the linear coefficient $\beta > 0$ represents the
consumers’ sensitivity to quality effort and is constant across all consumers. The expected
consumer utility, $U_{in} = a_i + \beta x_i - hW_i(\lambda_i, \mu_i, x_i) - p_i$, does not depend on $n$, i.e. the individual
consumer. Given the arrival rate, all consumers receive the same expected utility from service.
For convenience, we let $z_i = (p_i + b - a_i)$ denote the base “effective” out-of-pocket cost for
service to the consumer (when accounting for the base service value and the value of the outside
option).
Common results from competitive queueing models provide the following market clearing condition as an equilibrium solution to the consumer choice model when consumers are homogeneous. Consumers play a mixed strategy such that in equilibrium all consumers are indifferent between joining a local queue or taking the outside option (Chen and Wan 2005, Afanasyev and Mendelson 2010).

Incorporating our base assumption of consumer homogeneity, we rewrite this consumer choice equilibrium condition as

\[ a + \beta x - hW_i(\lambda_i, \mu_i) - p_i = b \]  

(2.3)

Proceeding by backward induction and writing \( z_i = (p_i + b - a_i) \), we can rewrite the provider’s optimization problem as follows:

\[
\max_{x_i \geq 0, \mu_i \geq 0} \pi_i = r_i \lambda_i - q_i x_i - c_i \mu_i - k 
\]

(2.4)

\[
\beta x_i - h \frac{h}{\mu_i - \delta x_i - \lambda_i} - z_i = 0 \quad \forall i = 1, 2 
\]

(2.5)

\[
\lambda_1 + \lambda_2 \leq \Lambda 
\]

(2.6)

\[
\lambda_i \geq 0 \quad \forall i = 1, 2 
\]

(2.7)

\[
\lambda_i < \mu_i - \delta x_i \quad \forall i = 1, 2 
\]

(2.8)

In the above problem statement, Equation (2.5) represents the consumer equilibrium constraint (or market clearing constraint); Equation (2.6) is a demand constraint which states that the sum of the arrival rates at each provider must be less than or equal to the total market po-
tential; Equation (2.7) is a straightforward non-negativity constraint on the arrival rate at each provider, and Equation (2.8) is the stationarity constraint required for stability of the queuing model.

We make some assumptions on the parameters to eliminate trivial cases for our model. First, we assume $r_i > c_i$ such that an entering provider will choose a non-zero service rate. We also assume $q_i + c_i \delta \geq 0$ to eliminate the case where the cost tradeoff for the quality effort is trivial and the provider pushes its quality effort to infinity.

To examine the impacts of competition we will compare the duopoly equilibrium with the monopoly equilibrium. These comparisons of monopolistic and oligopolistic environments seem reasonable for a health care market which, as we mentioned in the Introduction, is largely consolidated with many urban areas dominated by a few major health systems. For the monopoly case we drop the subscripts and optimize with respect to the service capacity, $\mu$, and the quality effort, $x$, to obtain the following proposition.

**Proposition 2.1.** Assuming $q + c\delta \geq 0$, if it is profitable for the monopolist to enter, the optimal quality and capacity for the monopolist are given by $x^* = \sqrt{\frac{ch}{\beta (q + c\delta)}} + \frac{\delta}{\beta}, \mu^* = \Lambda + \frac{\delta x}{\beta} + (2c\delta + q) \sqrt{\frac{h}{\beta (c\delta + q)}}, \lambda^* = \Lambda$

*Proof. All proofs in the appendix.*

The above result shows that if the monopolist can enter the market profitably, it will choose to serve the full market potential (average arrival rate of those requiring medical service from the population). The monopolist chooses a quality effort and service rate such that the effective service rate, $\tau = \mu - \delta x$, is equal to the arrival rate, $\lambda = \Lambda$, plus a non-negative capacity buffer. The capacity buffer is given as $\frac{1}{W}$, where $W$ denotes the optimal expected wait time and is
given by $W = \sqrt{\frac{\beta c}{h(q+c\delta)}}$.

We write the optimal profits for the monopoly as

$$\pi^*(\Lambda) = (r-c)\Lambda - \frac{z(c\delta+q)}{\beta} - 2\sqrt{\frac{ch(c\delta+q)}{\beta}} - k \quad (2.9)$$

We define social welfare, $SW$, as the sum of industry profits and consumer welfare. We can write the social welfare for the optimal monopolist decisions as

$$SW_{\text{mon}} = \pi^*(\Lambda) + \Lambda b = (r-c)\Lambda - \frac{z(c\delta+q)}{\beta} - 2\sqrt{\frac{ch(c\delta+q)}{\beta}} - k + \Lambda b \quad (2.10)$$

Now we transition to an asymmetric competitive duopoly market structure with two providers. Let $\pi^*_i(\Lambda)$ denote the optimal profits of Provider $i = 1, 2$ if operating as a single provider in the market. We can derive $\pi^*_i(\Lambda)$ from Equation (2.9) by making the appropriate substitutions. Note that $\pi^*_i(\Lambda)$ is linearly increasing in $\Lambda$. Let $\Lambda_i$ denote the breakeven arrival rate for Provider $i$ when it operates as a monopolist (the value of $\Lambda$ such that Provider $i = 1, 2$ may enter the market and earn non-negative profits). Assume without loss of generality that $\Lambda_1 \leq \Lambda_2$. This assumption implies that Provider 1 has a lower breakeven market arrival rate than Provider 2.

Using this notation we write the duopoly equilibrium as follows

**Theorem 2.1.** The duopoly equilibrium of the above optimization problem can be described by the following four scenarios:

- If $\Lambda < \Lambda_1$, then neither provider operates in the market.

- If $\Lambda_1 \leq \Lambda < \Lambda_2$, then Provider 1 operates as a monopoly with $\lambda_1 = \Lambda$. 

29
If $\Lambda_2 \leq \Lambda < \Lambda_1 + \Lambda_2$, then either provider can operate as monopoly, capturing the full market ($\lambda_i = \Lambda$).

If $\Lambda_1 + \Lambda_2 \leq \Lambda$ then either provider operates as a monopoly or both firms enter such that

- $\lambda_1 + \lambda_2 = \Lambda; \lambda_i \geq \Lambda_i$
- $x_i = \sqrt{\frac{c_i h}{\beta(q_i + c_i \delta_i)}} + \frac{z_i}{\beta} + \frac{\delta_i q_i}{\beta} \pm \frac{h}{\beta c_i h (c_i \delta_i + q_i)} \sqrt{\frac{h}{\beta c_i h (c_i \delta_i + q_i)}}$, $\lambda_i = \Lambda$
- $W_i = \sqrt{\frac{\beta c_i}{h(q_i + c_i \delta_i)}}$.

Comparing the duopoly equilibrium with the monopoly equilibrium, we obtain several key results regarding the impacts of competition. First, we observe that if any provider enters the market, the full market potential is captured by the entering provider(s), i.e. competition does not impact the total number of consumers served by the market because 1) consumers are homogenous and 2) any provider can increase its capacity to meet its demand. Second, the optimal quality decision at a given provider is unchanged regardless of whether the competitor enters. Though a competitor will claim some market share, a provider will choose the same quality effort under competition that it would choose with no competitor. Quality does not escalate at a given provider but will increase within the industry due to competition. Third, each provider adjusts its capacity to match its relative demand such that wait time remains the same at a given provider whether or not the competitor enters the market. From these results we see that competition does not impact quality or wait-time at a given provider. From the consumers’ point of view, there is no change in the overall service received as the quality and wait-time are not affected by competition. Whether there are one or two providers in the market, the consumers’ expected utility at any entering provider is equal to that of the outside option.

Competition does impact the overall industry expenditure on quality effort and capacity.
When two providers are in the market, both invest in quality and capacity, yet no additional consumers are served and the utility consumers receive from service is unchanged. When the firms have symmetric cost and price parameters, i.e. \( c_i = c, q_i = q, z_i = z, r_i = r \), then competition will lead to duplication but not escalation of investments in quality and capacity buffers. This supports the Medical Arms Race argument that overall industry expenditure on quality and capacity will increase, but it may happen via duplication as opposed to strictly escalation.

As with the monopolist case, we can write the equilibrium expected profits for Provider \( i = 1, 2 \) when capturing \( \lambda_i \) share of the market as:

\[
\pi^*_i(\lambda_i) = (r_i - c_i) \lambda_i - \frac{z_i (c_i \delta_i + q_i)}{\beta} - 2 \sqrt{\frac{c_i h (c_i \delta_i + q_i)}{\beta}} - k \tag{2.11}
\]

If only a single provider enters the market, we can write the social welfare as in the monopolist case above. When both providers enter the market, we write the social welfare as

\[
SW_{duo} = \pi^*_1(\lambda_1) + \pi^*_2(\lambda_2) + \Lambda b \tag{2.12}
\]

\[
= (r_1 - c_1) \lambda_1 - \frac{z_1 (c_1 \delta_1 + q_1)}{\beta} - 2 \sqrt{\frac{c_1 h (c_1 \delta_1 + q_1)}{\beta}} + k
\]

\[
+ (r_2 - c_2) \lambda_2 - \frac{z_2 (c_2 \delta_2 + q_2)}{\beta} - 2 \sqrt{\frac{c_2 h (c_2 \delta_2 + q_2)}{\beta}} + k + \Lambda b
\]

Note that the consumer surplus does not change because the full market is served and consumers receive the same utility regardless of competition due to wait-time effects. Let \( SW_i \) denote the social welfare when Provider \( i \) operates as a monopolist. We can write the change...
in social welfare due to simultaneous competition with Provider $j \neq i$ as

\[
SW_i - SW_{duo} = (r_i - c_i) \lambda_j - \pi_j^* (\lambda_j) \\
= \lambda_j ((r_i - c_i) - (r_j - c_j)) + \frac{z_j (c_j \delta_j + q_j)}{\beta} + 2 \sqrt{\frac{c_j h (c_j \delta_j + q_j)}{\beta}} + k \\
= \lambda_j ((r_i - c_i) - (r_j - c_j)) + \frac{c_j}{W_j^*} + q_j x_j^* + k
\]

Assume without loss of generality that $r_1 - c_1 \geq r_2 - c_2$ (i.e. Provider 1 is the high margin firm). The above equation states that a social welfare loss will always occur when both firms enter the market compared to when Provider 1 alone operates in the market. The reason is that the sum of the second and third terms must be positive for $x_j^* \geq 0$. However, the equation also shows that compared to a market with a low-margin monopolist, a duopoly market with an additional high-margin competitor may have higher social welfare. The implications for health care policy is that a monopoly market leads to higher social welfare unless adding an additional provider means adding a higher margin (i.e. lower capacity cost) provider, in which case competition may be beneficial in terms of social welfare. Adding a second provider at lower cost could happen in cases where a new provider has a disruptive innovation which allows it to enter the market at low cost, or the new provider is another major health system which benefits from even greater economies of scale and scope.

Note that Theorem 1 and the above interpretation does not apply to a market-entry game. When we refer to entry, we simply refer to any simultaneous equilibrium where both providers choose capacity strictly greater than zero, i.e. $\mu_i > 0 \ \forall i = 1, 2$. 
2.3.2 Special Case, $\delta = 0$

For convenience, we consider the symmetric case where $r_i = r$, $c_i = c$, $q_i = q$, and $p_i = p$ and $\delta = 0$ for the remainder of this section. The assumption that $\delta = 0$ simply implies there is no correlation between quality-effort and effective capacity; i.e. the provider can increase quality effort without slowing its service speed.

Now let $\Lambda_1 = \Lambda_2 = \Lambda_0$ where $\Lambda_0$ denotes the breakeven profit for the symmetric provider. From Theorem 1 we know that if $\Lambda < \Lambda_0$, then neither provider operates in the market, and if $\Lambda_0 \leq \Lambda < 2\Lambda_0$, then either Provider can operate as monopoly, capturing the full market ($\lambda_i = \Lambda$). If $2\Lambda_0 \leq \Lambda$ then either provider operates as a monopoly or both firms enter such that $\lambda_1 + \lambda_2 = \Lambda$; $\lambda_i \geq \Lambda_0$ and $x_i = \sqrt{\frac{chq}{\beta} + \frac{q}{\beta^2}}$, $\mu_i = \lambda_i + \sqrt{\frac{hq}{\beta c}}$. The equilibrium profits for the symmetric case are given by

$$\pi^*(\lambda_i) = (r - c)\lambda_i - 2\sqrt{\frac{chq}{\beta}} - \frac{qz}{\beta} - k$$  \hspace{1cm} (2.14)

The change in social welfare for the symmetric case can then be written as

$$SW_{mon} - SW_{duo} = 2\sqrt{\frac{chq}{\beta}} + \frac{qz}{\beta} + k$$  \hspace{1cm} (2.15)

Since the fixed operating cost is non-negative, i.e. $k > 0$, then there will be a loss in social welfare due to competition, because $2\sqrt{\frac{hcq}{\beta} + \frac{qz}{\beta^2}} = \frac{c}{\beta^2} + qx^* \geq 0$. 

33
Table 2.1: Comparative Statics: Optimal Quality, Wait Time, Profit, and Social Welfare

<table>
<thead>
<tr>
<th></th>
<th>(x_i^*)</th>
<th>(\mu_i^*)</th>
<th>(W_i^*)</th>
<th>(\pi_i^* (SW^*))</th>
<th>SW Loss (symmetric)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A)</td>
<td>no change</td>
<td>increases</td>
<td>no change</td>
<td>increases</td>
<td>no change</td>
</tr>
<tr>
<td>(r_i)</td>
<td>no change</td>
<td>no change</td>
<td>no change</td>
<td>increases</td>
<td>no change</td>
</tr>
<tr>
<td>(c_i)</td>
<td>increases</td>
<td>decreases</td>
<td>increases</td>
<td>decreases</td>
<td>increases</td>
</tr>
<tr>
<td>(q_i)</td>
<td>decreases</td>
<td>increases</td>
<td>decreases</td>
<td>decreases</td>
<td>increases</td>
</tr>
<tr>
<td>(h)</td>
<td>increases</td>
<td>increases</td>
<td>decreases</td>
<td>decreases</td>
<td>increases</td>
</tr>
<tr>
<td>(\beta)</td>
<td>not monotone</td>
<td>decreases</td>
<td>increases</td>
<td>increases</td>
<td>decreases</td>
</tr>
<tr>
<td>(z_i)</td>
<td>increases</td>
<td>increases if (\delta &gt; 0), decreases o.w.</td>
<td>no change</td>
<td>decreases</td>
<td>increases</td>
</tr>
<tr>
<td>(\delta_i)</td>
<td>decreases</td>
<td>increases</td>
<td>decreases</td>
<td>decreases</td>
<td>increases</td>
</tr>
</tbody>
</table>

2.3.3 Sensitivity Analysis

To gain further insight into how the optimal decisions behave with respect to the parameters, we perform comparative statics on the optimal decisions and key outputs of the model. Table 2.1 summarizes the monotonicity of the optimal quality effort, wait time, profit, and social welfare loss. Since the structure of the optimal solution is the same whether one or two firms compete in the market, the comparative statics are the same for both cases (except at the boundaries). We obtain several key insights into the providers’ optimal quality effort and capacity decisions from the table.

The first insight obtained is that in equilibrium providers balance the tradeoff between costs of quality effort and capacity. When the cost of capacity increases and all else remains equal, capacity becomes more expensive relative to quality effort, and the provider invests less in capacity and more in quality effort. The reverse is true for when the cost of quality effort increases. Given fixed prices, a provider will balance the costs of a quality effort increase and a wait time reduction in order to capture the respective market share. The managerial implication for health providers is that quality improvement and capacity expenditures should be balanced.
with one another relative to their costs. Intuitively, these tradeoffs can be thought of as moving along the respective edge of the Iron Triangle. This explains why a provider located in a downtown area where capacity expansion is limited, would spend more on quality efforts such as infrastructure or facility renovation. The policy implication is that higher costs lead to lower provider profits without impacting consumer welfare, therefore contributing to a greater loss of social welfare. To maximize social welfare, policy should continue to emphasize innovation which lowers marginal costs of capacity expansion and quality-improvement effort.

The second insight is that changes in either price or out-of-pocket payment do not impact wait-time, but quality effort increases with out-of-pocket price. According to Table 2.1, a change in the price does not affect the quality effort or wait time, it only affects the providers’ profits, the industry profit, and social welfare. On the other hand, if the out-of-pocket payment increases, thereby increasing $z_i$, consumers are paying more for service and the provider increases the quality effort accordingly to compensate for the utility lost, which leads to higher costs. Adjusting the out-of-pocket payment has no effect on optimal wait time. Higher prices generate higher profits for the provider but do not impact the quality effort or wait time seen by the consumer unless we assume the out-of-pocket payment is a function of the price. For example, if consumers pay $p_i = p_0 + \alpha r_i$ where $p_0$ is the fixed co-pay and $\alpha$ is the percent of the price paid by the consumer, then increasing the price would lead to an increase in quality effort, but no impact on the optimal wait time. Higher prices lead to higher social welfare because industry profit increases. Higher prices do not impact the loss of social welfare, unless a correlation exists between price and out-of-pocket payment. For example, if consumer out-of-pocket payment is structured as above, then higher prices lead to greater loss of social welfare. These results have implications for policy makers and analysts as they demonstrate how changes in price and out-
of-pocket made by third-party payers or policy makers may impact equilibrium quality effort and wait time. One might hypothesize that if out-of-pocket payments drop, moral hazard would increase and wait times would go up. In this model, however, quality effort is the lever used by the provider(s) to adjust to such a change in payments, as the provider would decrease its quality effort in order to leave wait times unchanged.

The third insight is that higher wait sensitivity leads to higher quality and lower wait, whereas higher quality sensitivity may lead to lower quality effort. From Table 2.1 we see that increasing wait time sensitivity leads to lower wait times and higher quality effort. The complement is not true with regard to quality effort sensitivity. Increasing quality effort sensitivity always leads to higher wait times because consumers are willing to wait longer for a service which gives them higher value. Increasing quality sensitivity does not necessarily lead to higher quality effort. While optimal quality effort $x^*$ is not always increasing in quality effort sensitivity, $\beta$, the optimal utility consumers earn from quality effort, $\beta x^*$, is increasing in $\beta$. Equilibrium quality effort at a given provider is increasing in quality effort sensitivity if and only if $\beta \leq \frac{4\epsilon^2(q+\epsilon\delta)}{ch}$ and decreasing otherwise. The result may seem counter-intuitive as it states that under certain conditions, namely when the consumer sensitivity to quality is low, an increase in consumer quality sensitivity may cause the provider to invest less in quality effort. If quality sensitivity increases, the consumers are willing to wait longer for service. If the optimal wait-time is relatively high, the provider would prefer to decrease quality effort so that consumers receive the same effective value from service and wait times do not increase too dramatically. Optimal (equilibrium) service rate at a given provider is increasing in quality effort sensitivity if and only if $\beta \leq \frac{4\epsilon\delta^2z^2(q+\epsilon\delta)}{h(3\epsilon\delta(q+\epsilon\delta)+q^2)}$ and increasing otherwise. The implication for managers and policy makers is that increasing quality sensitivity can increase wait times, and it is important to understand
the value consumers in the market receive from service relative to the disutility they receive from wait time.

From Table 2.1 we also find that the loss of social welfare due to competition is higher when costs are higher. The implication for policy makers is that programs, such as Certificate of Need, aimed at limiting competition in health care markets should continue to emphasize controlling high-cost medical services where the loss to society due to competition is greater, e.g. transplant services, cardiac surgery services, advanced radiation and imaging equipment. Social welfare loss is higher when the out-of-pocket payment is higher. The loss of social efficiency due to competition can be mitigated if government regulation or insurers reduce the out-of-pocket payment. However, this reduction in social welfare loss will be connected to a reduction in quality effort. Social welfare loss is also increasing in consumer sensitivity to wait-time and decreasing in consumer sensitivity to quality. Previous economic theory suggests that a loss of social welfare will occur when prices are fixed and firms compete on quality, but as discussed in the Literature Review section of this chapter, these papers have not explicitly incorporated wait-time dynamics. Our analysis shows that including consumer sensitivity to wait-time only increases the loss of social welfare as the loss is present even when the wait-time sensitivity, $h$, is zero. The loss of social welfare increases as wait-time sensitivity increases because a single provider can more efficiently invest in service capacity as a buffer against arrival and service variation in the market.

It is also important to understand the effect of the parameter $\delta$ which represents the correlation between quality effort and effective capacity. One key finding from these results is that the loss of social welfare is higher when $\delta$ is higher (higher $\delta$ implies a greater reduction in capacity). This result then agrees with the earlier intuition that higher costs imply a greater loss of social
welfare due to competition. When \( \delta > 0 \), higher out-of-pocket payment leads to higher capacity because in order to keep wait times the same, the provider must invest in additional capacity to offset the higher quality effort. When \( \delta < 0 \), then an increase in out-of-pocket payment leads to higher quality effort which increases the effective capacity and allows the firm to lower its investment in base capacity, \( \mu \).

2.4 Fixed Price Competition: Quality Effort Reduction

In the base model we find that provider competition does not impact the quality effort at a given provider, but leads to duplication in the market. In this section, we consider some variations on the base model where fixed-price competition between health providers may lead to lower quality effort at a given provider.

2.4.1 Utilization-Dependent Utility

One limitation of the base model is the simplifying assumption that service times and arrival times are exponential. In particular, these assumptions may over-estimate the variability in clinic capacity. To check for robustness, we derive results for a congestion-dependent utility function which does not incorporate expected wait-time into the utility function directly, but instead uses utilization-sensitivity. We let \( \rho = \frac{\lambda}{\mu - \delta x} \) represent the capacity utilization of the health system and \( h \) now represents the per unit disutility from higher utilization.

We adjust the provider’s optimization problem as follows:
\[
\max_{x_i \geq 0, \mu_i \geq 0} \pi_i = r_i \lambda_i - q_i x_i - c_i \mu_i - k
\]  \quad (2.16)

\[
\beta x_i - \frac{h \lambda_i}{\mu_i - \delta x_i} - z_i = 0 \quad \forall i = 1, 2
\]  \quad (2.17)

\[
\lambda_1 + \lambda_2 \leq \Lambda
\]  \quad (2.18)

\[
\lambda_i \geq 0 \quad \forall i = 1, 2
\]  \quad (2.19)

\[
\lambda_i < \mu_i - \delta x_i \quad \forall i = 1, 2
\]  \quad (2.20)

Compared to the base model, the primary difference is the second term in Equation (2.17) which now references utilization as opposed to wait-time. We will assume \(c\delta + q > 0\) to eliminate trivial solutions and also assume \(\Lambda \leq \frac{h(c\delta + q)}{\beta c}\) so that the optimal solution which covers the full market satisfies the stability condition. Solving for the monopoly outcome, we obtain the following result:

**Proposition 2.2.** If profitable to enter the market, the monopoly solution to the above problem is given by

\[
x = \frac{z}{\beta} + \sqrt{\frac{c h \Lambda}{\beta (c \delta + q)}}; \quad \mu = \frac{\delta z}{\beta} + (q + 2c\delta) \sqrt{\frac{h \Lambda}{c \beta (c \delta + q)}}; \quad \lambda = \Lambda; \quad \rho = \sqrt{\frac{c \delta \Lambda}{h (c \delta + q)}}; \quad \pi = r \Lambda - \frac{z (c \delta + q)}{\beta} - 2 \sqrt{\frac{c h \Lambda (c \delta + q)}{\beta}} - k.
\]

Comparing this solution with the optimal monopoly solution from the base case we see several similarities. In both models, the monopoly provider captures the full-market (if any consumers at all). In neither model does the price, either full price or out-of-pocket payment, impact the optimal wait-time (or congestion). There are also important differences in the results of this model and the base model. In this model, the market potential has a concave, increasing impact on both the quality effort and the capacity decisions as opposed to only a linearly increasing impact on capacity in the base model. This will affect our base case results regarding the impact
of competition on quality, access, and social welfare as stated in the following Theorem. Let $\Lambda_0$ denote the breakeven arrival rate for the monopoly.

**Theorem 2.2.** The duopoly equilibrium to the congestion-dependent model is given by the following three scenarios:

If $\Lambda < \Lambda_0$, then neither provider operates in the market.

If $\Lambda_0 \leq \Lambda < 2\Lambda_0$ then either provider can operate as monopoly, capturing the full market ($\lambda_i = \Lambda$).

If $2\Lambda_0 \leq \Lambda$ then either provider operates as a monopoly or both firms enter such that

$$\lambda_1 + \lambda_2 = \Lambda; \lambda_i \geq \Lambda_0$$

$$x_i = \frac{z}{\beta} + \sqrt{\frac{ch\lambda_i}{\beta(c\delta + q)}}; \mu_i = \frac{\delta z}{\beta} + (q + 2c\delta) \sqrt{\frac{h\lambda_i}{c\beta(c\delta + q)}}; \rho_i = \sqrt{\frac{c\beta\lambda_i}{h(c\delta + q)}}.$$  

The symmetric duopoly equilibrium structure is similar to the base model except that now the quality and access at each provider will be lower than that at the monopoly provider.

The difference in total quality effort in the market is given by

$$x_{mon} - (x_1 + x_2) = -\frac{z}{\beta} + \sqrt{\frac{ch}{\beta (c\delta + q)}} \left( \sqrt{\Lambda} - \sqrt{\lambda_1} - \sqrt{\lambda_2} \right)$$  

(2.21)

If $z \geq 0$, then competition will imply a higher quality effort in the market as both terms in Equation (2.21) will be non-positive and the second term will be strictly negative.

The social welfare in the duopoly setting is given by

$$SW_{DUO} = r\Lambda - \frac{2z (c\delta + q)}{\beta} - 2 \sqrt{\frac{ch\lambda_2 (c\delta + q)}{\beta}} - 2 \sqrt{\frac{ch\lambda_1 (c\delta + q)}{\beta}} - 2k$$  

(2.22)
In this model setting, the symmetric market split, $\lambda_i = \Lambda/2$, minimizes the social welfare and maximizes total quality effort in the market.

The non-linearity with respect to the market potential has additional implications regarding the impact of competition. If both providers enter, the quality and wait time at each provider will be lower than in the monopoly equilibrium. The change in social welfare between the monopoly and the symmetric duopoly equilibrium when Provider 1 captures $\lambda_1$ will be given by

$$SW_{Mon} - SW_{Duo} = \frac{z(c\delta + q)}{\beta} + k + 2\sqrt{ch(c\delta + q)} \left(\sqrt{\lambda_1} + \sqrt{\Lambda - \lambda_1} - \sqrt{\Lambda}\right)$$ (2.23)

If $z \geq 0$, then we are guaranteed there will be a loss of social welfare due to competition as all three terms in Equation (2.23) will be non-negative. However, if $z < 0$, social welfare may increase or decrease. The actual loss of social welfare will depend on how the market is split between the competing providers. The maximum loss of social welfare will be at the symmetric split of $\lambda_i = \Lambda/2$ and the minimum loss of social welfare is obtained by giving as much demand as possible to one provider $\lambda_i = \Lambda_0$ and $\lambda_j = \Lambda - \Lambda_0$ for $i = 1, 2; j \neq i$.

2.4.2 Technical vs. Functional Quality

In our base model we assume that quality effort may impact the value the consumer receives from service and/or the effective provider capacity. However, quality effort may also impact the service outcome in terms of probability of a successful service outcome which does not require repeat service. We assume that this probability of a successful outcome is given by $\gamma x_i$ where $\gamma > 0$ represents the correlation between quality effort and probability of the intended outcome.
In a healthcare context, we can think of the probability of an unsuccessful service outcome as representing rework, readmission, follow-up due to medical error, or a try-and-see approach to diagnosis and treatment, etc. We assume that the provider does not earn additional revenue for the rework, but the results are qualitatively similar when revenue is earned for each patient visit (when there are multiple patient encounters). We write the provider optimization problem as follows.

\[
\max_{x_i \geq 0, \mu_i \geq 0} \pi_i = r_i \lambda_i - q_i x_i - c_i \mu_i - k
\]  

(2.24)

\[
\beta x_i - \frac{h}{\mu_i - \frac{\lambda_i}{\gamma x_i}} - z_i = 0 \quad \forall i = 1, 2
\]  

(2.25)

\[
\lambda_1 + \lambda_2 \leq \Lambda
\]  

(2.26)

\[
\lambda_i \geq 0 \quad \forall i = 1, 2
\]  

(2.27)

\[
\frac{\lambda_i}{\gamma x_i} < \mu_i \quad \forall i = 1, 2
\]  

(2.28)

\[
\gamma x_i \leq 1
\]  

(2.29)

In optimality, the monopoly solution is to take the full market and the solution satisfies the following equations.

\[
\frac{\Lambda}{x^2} + \frac{\beta h \gamma}{(\beta x - z)^2} = \frac{q \gamma}{c}
\]  

(2.30)

\[
\mu = \frac{\Lambda}{\gamma x} + \frac{h}{\beta x - z}
\]  

(2.31)
To simplify these conditions and obtain a closed-form solution we will consider the case where \( \beta = 0 \), i.e. quality effort only impacts the probability of a successful service outcome and not the service valuation. We will assume \( z < 0 \) so that the optimal solution is feasible according to the stability condition. We also assume \( \gamma \) is small enough such that \( \gamma x \leq 1 \)

**Proposition 2.3.** The monopoly solution to the above problem is given by

\[
x = \sqrt{\frac{c\Lambda}{\gamma}}, \quad \mu = \sqrt{\frac{\Lambda}{c\gamma}} - \frac{h}{2}; \quad \lambda = \Lambda; \quad W = -\frac{z}{h}; \quad \pi = r\Lambda + \frac{hc}{2} - 2\sqrt{\frac{2c\Lambda}{\gamma}} - k.
\]

Comparing this special case solution with the optimal solution from the base case we see several similarities. In both models, the optimal solution for the monopoly is to capture the full-market. In neither model does the (reimbursement) price or out-of-pocket payment, impact the optimal wait-time (or congestion).

There are also differences in the results of this model and the base model. In this model, the optimal wait time decreases in the out-of-pocket price but the out-of-pocket price does not impact the quality effort. This supports the idea of moral hazard, that as consumers pay less, wait times will increase as more consumers try to join the queue. Similar to the congestion-dependent utility model, the market potential has a concave, increasing impact on both the quality effort and the capacity decisions as opposed to only a linearly increasing impact on capacity in the base model. Letting \( \Lambda_0 \) be the break-even arrival rate, the structure of the duopoly equilibrium follows as in the base model.

**Theorem 2.3.** The duopoly equilibrium to the congestion-dependent model is given by the following three scenarios:

If \( \Lambda < \Lambda_0 \), then neither provider operates in the market.

If \( \Lambda_0 \leq \Lambda < 2\Lambda_0 \) then either provider can operate as monopoly, capturing the full market
If $2\Lambda_0 \leq \Lambda$ then either provider operates as a monopoly or both firms enter such that

$$\lambda_1 + \lambda_2 = \Lambda; \lambda_i \geq \Lambda_0$$

$$x_i = \sqrt{\frac{c\lambda_i}{q^2}}; \mu_i = \sqrt{\frac{\lambda_i}{c^2}} - \frac{h}{\varepsilon}; W_i = -\frac{x_i}{h}.$$

The non-linearity with respect to the market potential has additional implications regarding the impact of competition. If both providers enter, the quality and wait time at each provider will be lower than in the monopoly equilibrium. The symmetric duopoly equilibrium structure is similar to the base model except that now the quality and access at each provider will be lower than that at the monopoly provider.

The difference in total quality effort is given by

$$x_{mon} - (x_1 + x_2) = \sqrt{\frac{c}{q^2}} \left( \sqrt{\Lambda} - \sqrt{\lambda_1} - \sqrt{\lambda_2} \right)$$

(2.32)

Since $\left( \sqrt{\Lambda} - \sqrt{\lambda_1} - \sqrt{\lambda_2} \right)$ is always negative, the total quality effort in the duopoly market will be more than in the monopoly market. The change in social welfare between the monopoly and the symmetric duopoly equilibrium when Provider 1 captures $\lambda_1$ will be given by

$$SW_{Mon} - SW_{Duo} = -\frac{hc}{\varepsilon} + k + 2\sqrt{qc} \left( \sqrt{\lambda_1} + \sqrt{\Lambda - \lambda_1} - \sqrt{\Lambda} \right)$$

(2.33)

Since we assume $z < 0$, social welfare must decrease due to competition as all terms in Equation (2.33) must be non-negative.
2.5 Fixed-Price Competition: Quality Effort Escalation

2.5.1 Quality cost per arrival

Earlier we discussed the limitation of a linear cost function. One alternative we consider is a cost function where the margin earned per arrival decreases linearly in the quality effort, e.g. higher quality may imply additional tests, safety checks, sanitation steps, etc. associated with each individual arrival. In this setting, the cost of quality effort is incurred per arriving consumer as opposed to a one-time cost per period. We write the revised optimization problem as follows;

\[
\begin{align*}
\max_{x_i \geq 0, \mu_i \geq 0} \pi_i &= (r_i - qx_i) \lambda_i - c_i \mu_i - k \\
\beta x_i - \frac{h}{\mu_i - \lambda_i} - z_i &= 0 \quad \forall i = 1, 2 \\
\lambda_1 + \lambda_2 &\leq \Lambda \\
\lambda_i &\geq 0 \quad \forall i = 1, 2 \\
\lambda_i &< \mu_i - \delta x_i \quad \forall i = 1, 2
\end{align*}
\]

The primary difference between this model and the base model is in the objective function where now the cost of quality effort \( qx \) is incurred per arrival. Solving the above optimization problem for the monopolist, we obtain the following proposition

**Proposition 2.4.** The monopoly solution to the above problem is given by

\[
\begin{align*}
\mu &= \Lambda + \frac{\delta z}{\beta} + (q\Lambda + 2c\delta) \sqrt{\frac{h}{\beta c^2 + q\Lambda}}; \\
\lambda &= \Lambda; \\
W &= \sqrt{\frac{\beta c}{h(c^2 + q\Lambda)}}, \\
\pi &= r\Lambda - \frac{z(c^2 + q\Lambda)}{\beta} - 2\sqrt{\frac{ch(c^2 + q\Lambda)}{\beta}} - k.
\end{align*}
\]

This model provides results which are similar to the base case in that the optimal quality
effort, capacity, and wait time take nearly identical forms as in the base case. However, now the
total market arrival rate will impact all three in a non-linear fashion. Compared with previous
extensions, the quality effort and wait-time in this extension of the model are decreasing in the
market arrival rate. Competition strictly increases the quality effort at each provider but in
return, the wait time will be higher.

**Theorem 2.4.** The duopoly equilibrium when quality costs are incurred per arrival is given by
the following three scenarios:

If \( \Lambda < \Lambda_0 \), then neither provider operates in the market.

If \( \Lambda_0 \leq \Lambda < 2\Lambda_0 \) then either provider can operate as monopoly, capturing the full market
\( (\lambda_i = \Lambda) \).

If \( 2\Lambda_0 \leq \Lambda \) then either provider operates as a monopoly or both firms enter such that

\[
\lambda_1 + \lambda_2 = \Lambda; \lambda_i \geq \Lambda_0
\]

\[
x_i = \frac{\lambda_i}{\beta} + \sqrt{\frac{c \lambda_i}{\beta (\sigma + q \lambda_i)}}, \quad \mu_i = \lambda_i + \frac{\lambda_i}{\beta} + (q \lambda_i + 2c \delta) \sqrt{\frac{h}{\beta c (\sigma + q \lambda_i)}},
\]

\[
W_i = \sqrt{\frac{\beta c}{h (\sigma + q \lambda_i)}}.
\]

The non-linearity with respect to the market potential has additional implications regarding
the impact of competition. If both providers enter, the quality and wait time at each provider
will be higher than in the monopoly equilibrium. The structure of the symmetric duopoly
equilibrium is similar to that in the base case, but the impact of competition on quality and
access is different; now the quality and access at each provider will be strictly higher than that
at the monopoly provider.

The difference in total quality effort is given by
\[ x_{mon} - (x_1 + x_2) = -\frac{z}{\beta} + \sqrt{\frac{ch}{\beta (c\delta + q)}} \left( \sqrt{\frac{1}{\Lambda}} - \sqrt{\frac{1}{\lambda_1}} - \sqrt{\frac{1}{\lambda_2}} \right) \] (2.39)

The change in social welfare between the monopoly and the symmetric duopoly equilibrium will be given by

\[ SW_{Mon} - SW_{Duo} = \frac{z (q + c\delta)}{\beta} + k + \sqrt{\frac{2ch (c\delta + q)}{\beta}} \left( \sqrt{\lambda_1} + \sqrt{\lambda_2} - \sqrt{\Lambda} \right) \] (2.40)

If \( z \geq 0 \), then we are guaranteed there will be a loss of social welfare due to competition as all three terms in Equation (2.40) will be non-negative. However, if \( z < 0 \), social welfare may increase or decrease. The actual loss of social welfare will depend on how the market is split between the competing providers.

### 2.5.2 Capacity-Dependent Cost of Quality

An alternative method for modeling a volume-based quality cost is to consider a cost function where the cost of quality effort depends on the capacity, e.g. it may cost more to conduct additional noise abatement at a larger facility than at a smaller facility. In this setting, the cost of quality effort is incurred per unit of capacity as opposed to only being linked to the quality effort. We write the revised optimization problem as follows;
The primary difference between this model and the base model is in the objective function where now the cost of quality effort $qx$ is incurred per unit of capacity. Solving the above optimization problem for the monopolist, we obtain the following proposition

**Proposition 2.5.** The monopoly solution to the above problem is given by

$$x = \frac{\pi}{\beta} + \sqrt{\frac{(zq+c\beta)}{2\beta q\Lambda}};$$

$$\mu = \Lambda + \sqrt{\frac{2hq\Lambda}{(zq+c\beta)}};$$

$$\pi = \Lambda (r - c) - \frac{q(2z\Lambda + h)}{\beta} - \frac{2}{\beta} \sqrt{2hq\Lambda (zq + c\beta) - k}.$$
If $\Lambda_0 \leq \Lambda < 2\Lambda_0$ then either provider can operate as monopoly, capturing the full market ($\lambda_i = \Lambda$).

If $2\Lambda_0 \leq \Lambda$ then either provider operates as a monopoly or both firms enter such that

$$\lambda_1 + \lambda_2 = \Lambda; \lambda_i \geq \Lambda_0$$

$$x_i = \frac{z}{\beta} + \sqrt{\frac{(zq+c\delta)}{2\beta q\lambda_i}}; \mu_i = \lambda_i + \sqrt{\frac{2hq\lambda_i}{(zq+c\beta)}}; W_i = \sqrt{\frac{(zq+c\beta)}{2\lambda_i qh}}.$$  

The non-linearity with respect to the market potential has additional implications regarding the impact of competition. If both providers enter, the quality and wait time at each provider will be higher than in the monopoly equilibrium. The structure of the symmetric duopoly equilibrium is similar to that in the base case, but the impact of competition on quality and access is different; now the quality and access at each provider will be strictly higher than that at the monopoly provider.

The difference in total quality effort in the market is given by

$$x_{\text{mon}} - (x_1 + x_2) = \frac{z}{\beta} + \sqrt{\frac{ch}{\beta (c\delta + q)}} \left( \sqrt{\frac{1}{\lambda_1}} + \sqrt{\frac{1}{\lambda_2}} - \sqrt{\frac{1}{\Lambda}} \right)$$  

The change in social welfare between the monopoly and the symmetric duopoly equilibrium when Provider 1 captures $\lambda_1$ will be given by

$$SW_{\text{Mon}} - SW_{\text{Duo}} = \frac{z (q + c\delta)}{\beta} + k + \sqrt{\frac{2ch (c\delta + q)}{\beta}} \left( \sqrt{\lambda_1} + \sqrt{\lambda_2} - \sqrt{\Lambda} \right)$$  

If $z \geq 0$, then we are guaranteed there will be a loss of social welfare due to competition as
all three terms in Equation (2.47) will be non-negative. However, if \( z < 0 \), social welfare may increase or decrease. The actual loss of social welfare will depend on how the market is split between the competing providers.

### 2.6 Single Decision Variable Model

In this sub-section we consider a variation on our model where quality and service capacity decisions are linked by a single decision variable. This variation applies to cases where the provider is constrained in terms of its capacity or quality effort. Relevant applications of this model in health care services include cases where the provider is unable to greatly expand its bottleneck resources, e.g. facility space, limited number of diagnostic imaging machines, limited supply of certain specialists.

Assume each provider chooses its quality effort \( x_i \) at a cost of \( q_i \) per unit, and that quality effort decreases effective capacity by \( \delta_i \) per unit and increases customer utility by \( \beta \) per unit. Let \( \mu_i \) equal the base capacity available to Provider \( i = 1, 2 \) when \( x_i = 0 \). We can write the duopoly optimization problem as follows:

\[
\begin{align*}
\max_{x_i \geq 0} \pi_i &= r_i \lambda_i - q_i x_i - k \\
\beta x_i - \frac{h}{\mu_i - \delta_i x_i - \lambda_i} - z_i &\geq 0 \quad \forall i = 1, 2 \quad (2.48) \\
\lambda_1 + \lambda_2 &\leq \Lambda \quad (2.50) \\
\lambda_i &\geq 0 \quad \forall i = 1, 2 \quad (2.51) \\
\lambda_i &< \mu_i - \delta x_i \quad \forall i = 1, 2 \quad (2.52)
\end{align*}
\]
We assume that \( q + r\delta > 0 \) to eliminate trivial cases. Further, we assume the parameters are such that \( x^* \geq 0; \lambda^* \geq 0; \) and \( \pi^* \geq 0, \) i.e. if the market potential were unlimited the monopolist could enter the market and earn non-negative profits by setting a non-negative quality effort.

Note that the consumer utility constraint in the above problem has two roots when the market constraint is binding. We will assume that the parameters are such that these roots exist and that they are non-negative, i.e. the minimum quality effort needed to capture the full-market potential is non-negative. This simplifying assumption eliminates some trivial boundary cases.

\[
x(\Lambda) = \frac{1}{2\beta\delta} \left( \mu \beta - \Lambda \beta + \delta z \pm \sqrt{(\mu \beta - \Lambda \beta - \delta z)^2 - 4\delta \beta h} \right) \geq 0
\]

Under these assumptions, the optimal solution to the monopolist problem is given by the following proposition.

**Proposition 2.6.** The optimal decision for the monopolist depends on the following two cases

- If \( q + r\delta \leq 0, \) then it is always optimal for the monopolist to capture the full market, i.e. \( \lambda = \Lambda \) and \( x = x(\Lambda) \) where \( x(\Lambda) \) is the minimum of the two solutions to \( \beta x - \frac{h}{\mu + \delta x - \Lambda} - z = 0. \)

- If \( q + r\delta > 0, \) then the monopolist sets \( x = \min \left( x(\Lambda), \sqrt{\frac{rh}{\beta(q + r\delta)}} + \frac{z}{\beta} \right) \) such that \( \lambda = \min \left( \mu - \frac{\delta z}{\beta} - (2r\delta + q) \sqrt{\frac{h}{r\beta(q + r\delta)}}, \Lambda \right). \)

If the monopolist chooses to capture the full market, the equilibrium profit and social welfare are

\[
\pi^*(\Lambda) = r\Lambda - qx(\Lambda) - k \quad (2.53)
\]
Unlike the base model, in some cases here the monopoly provider may choose not to capture
the entire market. Even if some consumers are un-served, the overall consumer surplus in
unchanged as consumers will play an equilibrium strategy such that the utility of the local
provider is equal to that of the outside option. Regardless, the total consumer surplus is still
given by $\Lambda_0 b$. Using this equilibrium result we can show that when prices and capacity are fixed,
there exists a region where the duopoly providers both enter the market and invest strictly less
in quality effort than the monopolist. If the demand constraint is binding, then wait times
will be lower than for the monopoly. If the monopolist chooses to capture the full market, the
equilibrium profit and social welfare are

$$\pi^* (\Lambda) = r\Lambda - qx (\Lambda) - k \quad (2.54)$$

$$SW_{\text{mon}} (\Lambda) = r\Lambda - qx (\Lambda) - k + \Lambda b \quad (2.55)$$

If the monopolist chooses not to capture the full market, we write the equilibrium profit and
social welfare as

$$\pi^* = r\mu - (q + r\delta) \left( \frac{z}{\beta} \right) - 2 \sqrt{\frac{rh (q + r\delta)}{\beta}} - k \quad (2.56)$$

$$SW_{\text{mon}} = r\mu - (q + r\delta) \left( \frac{z}{\beta} \right) - 2 \sqrt{\frac{rh (q + r\delta)}{\beta}} - k + \Lambda b \quad (2.57)$$

Let $\Lambda_0$ be the minimum market potential where the monopolist obtains breakeven profits.
Let $\Lambda_{\text{MAX}} = \mu - \frac{\delta z}{\beta} - (2r\delta + q) \sqrt{\frac{h}{r\beta(q+r\delta)}}$ be the arrival rate chosen by the monopolist if the total market potential were unbounded. Assume that at the breakeven rate, $\Lambda_0$, the total market arrival rate is binding for the monopolist, i.e. $\Lambda_0 \leq \Lambda_{\text{MAX}}$. Solving the symmetric duopoly problem we obtain the following proposition which captures the duopoly equilibrium.

**Theorem 2.6.** The duopoly Nash equilibrium of the symmetric single-decision model can be described by the following three scenarios:

- If $\Lambda < \Lambda_0$, then neither provider operates in the market.
- If $\Lambda_0 \leq \Lambda < \Lambda_0 + \Lambda_\infty$ then either Provider can operate as monopoly

$$x_i = \min \left( x(\Lambda), \sqrt{\frac{rh}{\beta(q+r\delta)} + \frac{z}{\beta}} \right).$$

- If $2\Lambda_0 \leq \Lambda < 2\Lambda_{\text{MAX}}$ then a continuum of equilibria exist where both providers enter and capture the full market

$$x_i = \min \left( x(\lambda_i), \sqrt{\frac{rh}{\beta(q+r\delta)} + \frac{z}{\beta}} \right)$$

$$\lambda_1 + \lambda_2 = \Lambda$$

$$\lambda_i \geq \Lambda_0.$$ 

- If $2\Lambda_{\text{MAX}} \leq \Lambda$, then a single equilibrium exists

$$x_i = \sqrt{\frac{rh}{\beta(q+r\delta)} + \frac{z}{\beta}}$$

$$\lambda_i = \Lambda_\infty.$$ 

Observe that there exists a region, $2\Lambda_0 \leq \Lambda < \Lambda_0 + \Lambda_{\text{MAX}}$, where it is possible to have either both providers enter, or one provider operate as a monopolist and capture a large enough market share to bar entry of the competitor. Now we write the symmetric provider profit and social welfare for the two cases where both providers enter the market and either 1) capture the
full market, or 2) capture part of the market.

If both providers enter and capture the full market, i.e. $2\Lambda_0 \leq \Lambda < 2\Lambda_{MAX}$, then we write the equilibrium profit and social welfare as

$$
\pi^*_i (\Lambda) = r\lambda_i - qx (\lambda_i) - k 
$$

(2.58)

$$
SW_{duo} = r\Lambda - q (x (\lambda_1) + x (\lambda_2)) - 2k + \Lambda b 
$$

(2.59)

If both providers enter and capture part of the market, i.e. $2\Lambda_{MAX} \leq \Lambda$, then we write the equilibrium profit and social welfare as

$$
\pi^*_i (\Lambda) = r\mu - (q + r\delta) \left( \frac{z}{\beta} \right) - 2 \sqrt{\frac{rh (q + r\delta)}{\beta}} - k 
$$

(2.60)

$$
SW_{duo} = 2r\mu - 2 (q + r\delta) \left( \frac{z}{\beta} \right) - 4 \sqrt{\frac{rh (q + r\delta)}{\beta}} - 2k + \Lambda b 
$$

(2.61)

According to Theorem 2.6, we can now compare the equilibrium results when there are one or two potential providers in the market. Clearly, if $\Lambda < \Lambda_0$ no provider may enter the market and the equilibrium results do not differ. If $\Lambda_0 \leq \Lambda < 2\Lambda_0$ then under duopoly competition only a single provider (one of the two) will operate. Therefore, once again the equilibrium results do not differ between the case of one provider or two providers. The interesting cases for comparison are the three cases where both firms may enter under a duopoly setting.

If the market potential is small such that the both the monopoly and the duopoly capture
the full market, i.e. \( 2\Lambda_0 \leq \Lambda < \Lambda_{\infty} \), then the difference in social welfare is given by

\[
SW_{mon} - SW_{duo} = q(x(\lambda_1) + x(\lambda_2) - x(\Lambda)) + k \tag{2.62}
\]

If the market potential is large enough that the monopoly leaves some consumers un-served but small enough that the duopoly captures the full market, i.e. \( 2\Lambda_0 \leq \Lambda < 2\Lambda_{\text{MAX}} \) and \( \Lambda_{\text{MAX}} \leq \Lambda < 2\Lambda_{\text{MAX}} \), then the difference in social welfare is given by

\[
SW_{mon} - SW_{duo} = r(\mu - \Lambda) - q(x(\lambda_1) + x(\lambda_2)) - (q + r\delta) \left( \frac{z}{\beta} \right) - 2\sqrt{r h (q + r\delta) \frac{\beta}{\beta}} + k \tag{2.63}
\]

If the market potential is large enough, the two duopoly providers operate as if monopoly providers and some consumers remain un-served, i.e. \( 2\Lambda_0 \leq \Lambda < 2\Lambda_{\text{MAX}} \). The difference in social welfare is given by

\[
SW_{mon} - SW_{duo} = -r\mu + (q + r\delta) \left( \frac{v}{\beta} \right) + 2\sqrt{r h (q + r\delta) \frac{\beta}{\beta}} + k \leq 0 \tag{2.64}
\]

In Equation (2.64), the market benefits from having a second provider because consumer surplus is unaffected due to the outside option, but industry profits increase because a second provider enters the market profitably due to a large segment which was previously un-served.
2.6.1 Special Case: $\delta = 0$

When $\delta = 0$, we have simplified the problem to optimizing the firm’s quality decision when capacity is fixed. Under this special case $x(\Lambda) = \frac{z}{\beta} + \frac{h}{\mu - \Lambda}$. According to Proposition 2.6 the optimal decision for the monopolist is given by

$$x = \min \left( \frac{z}{\beta} + \frac{h}{\beta (\mu - \Lambda)}, \frac{z}{\beta} + \sqrt{\frac{rh}{\beta q}} \right)$$

(2.65)

The optimal arrival rate is given by

$$\lambda = \min \left( \Lambda, \mu - \sqrt{\frac{hq}{\beta r}} \right)$$

(2.66)

For the special case $\delta = 0$, when both providers enter for the symmetric duopoly, the equilibrium is given by

$$x_i = \min \left( \frac{z}{\beta} + \frac{h}{\beta (\mu - \Lambda + \lambda_j)}, \frac{z}{\beta} + \sqrt{\frac{rh}{\beta q}} \right) \lambda_i + \lambda_j \leq \Lambda; \lambda_i \geq \Lambda_0$$

(2.67)

The optimal arrival rates are given by

$$\lambda_i = \min \left( \Lambda - \lambda_j, \mu - \sqrt{\frac{hq}{\beta r}} \right) \forall i = 1, 2$$

(2.68)

When the duopoly captures the full-market the social welfare can now be expressed as

$$SW_{duo} = r\Lambda - \frac{2qz}{\beta} - \frac{qh}{\beta} \left( \frac{1}{\mu - \lambda_1} + \frac{1}{\mu - \Lambda + \lambda_1} \right) + \Lambda b$$

(2.69)
Using the above function, we see that the social welfare for the duopoly will be maximized at \( \lambda_1 = \Lambda/2 \), assuming the queueing stability condition, \( \mu > \Lambda/2 \), is met. To maximize social welfare, the market potential should be split evenly between the two providers. This maximum social welfare is given by

\[
SW_{\text{duo}}^* = r\Lambda - \frac{2qz}{\beta} - \frac{2qh}{\beta} \left( \frac{1}{\mu - \Lambda/2} \right) + \Lambda b \tag{2.70}
\]

If both the monopolist and the duopoly capture the full market, the difference between the social welfare for the monopolist and the maximum duopoly social welfare is given by

\[
SW_{\text{mon}} - SW_{\text{duo}}^* = qz \frac{\beta}{\beta} + k + \frac{qh}{\beta} \left( \frac{2\mu - 3\Lambda}{(\mu - \Lambda)(2\mu - \Lambda)} \right) \tag{2.71}
\]

Solving for when the right hand side of Equation (2.71) is non-negative, we derive the following condition for the social welfare being maximized by having a single provider. Otherwise, social welfare is higher with two providers.

\[
\frac{k\beta}{hq} + \frac{z}{h} > \frac{-2\mu + 3\Lambda}{(\mu - \Lambda)(2\mu - \Lambda)} \tag{2.72}
\]

This is an important and interesting result because it shows that even when the number of consumers served by the local market is the same, two providers may be able to increase social welfare by providing lower quality effort and lower wait times. This contradicts previous economic theory and the Medical Arms Race argument which claims that fixed-price competition always leads to duplication/escalation of quality effort, generating higher costs, and lower social welfare.
2.6.2 Interpretation and Sensitivity Analysis

Competition may lead to increased social welfare when base capacity is fixed. We will use the special case of $\delta = 0$, in which capacity is fixed and the provider chooses only its quality effort, to intuitively explain how this may happen. Figure 2.1 provides an illustration of two numerical cases, showing how industry profit depends on market size for symmetric providers. Since consumer surplus does not vary across the models, the figure also provides insight regarding social welfare. The graphs show industry profit as a function of the total market potential. In the first graph, when $r = 100$, if the total market potential is less than 32 then neither provider can enter. First we look at the monopolist. If the market potential is at least 32, the monopolist captures the full market. If the market potential is at least 44, then the marginal cost of obtaining the additional consumer is too high, and the monopolist chooses to serve only part of the market. Since the monopolist is not receiving reimbursement for these un-served consumers, industry profits flatten. For the duopoly, if market potential is less than 64, then only one provider may enter and it behaves as the monopolist. If the market potential is at least 64 but less than 88, then both providers may enter as a duopoly and quality and wait-time at each duopoly provider are lower than in a monopoly market. Beyond a market potential of 88, the duopoly does not capture the full market and some consumers take the outside option. At a market potential of 73.895, the capacity gained from having a second provider in the market increases industry profits and thereby benefits social welfare as there is no difference in consumer welfare.

The second graph is similar except the reimbursement for services is higher. This means the providers can break-even with fewer consumers served. Providers are also willing to take on
more consumers before leaving some of the market un-served. First we look at the monopolist. If the market potential is at least 13.987, the monopolist captures the full market. If the market potential is at least 45.333, the monopolist chooses to serve only part of the market. For the duopoly, if market potential is less than 27.974, then only one provider may enter and it behaves as the monopolist. If the market potential is at least 27.974, but less than 90.666, then both providers may enter as a duopoly and quality and wait-time at each duopoly provider are lower than if a single monopoly operates in the market. Beyond a market potential of 90.666, the duopoly does not capture the full market and some consumers take the outside option. At a market potential of 57.177, the capacity gained from having a second provider in the market increases industry profits and social welfare. The two graphs illustrate how in one case it may be unprofitable for a second provider to enter even when part of the market is un-served. This is the case when $r = 100$ but not the case when $r = 225$. 
Table 2.2: Comparative Statics: Single Decision Model

<table>
<thead>
<tr>
<th></th>
<th>$x_i^*$ full</th>
<th>$x_i^*$ partial</th>
<th>$\lambda_i^*$</th>
<th>$W_i^*$</th>
<th>$\pi_i^* (SW^*)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$r_i$</td>
<td>no effect</td>
<td>increases</td>
<td>increases</td>
<td>increases</td>
<td>increases</td>
</tr>
<tr>
<td>$\Lambda$</td>
<td>increases</td>
<td>no effect</td>
<td>increases</td>
<td>no effect</td>
<td>increases</td>
</tr>
<tr>
<td>$q_i$</td>
<td>no effect</td>
<td>decreases</td>
<td>decreases</td>
<td>decreases</td>
<td>decreases</td>
</tr>
<tr>
<td>$h$</td>
<td>increases</td>
<td>increases</td>
<td>decreases</td>
<td>decreases</td>
<td>decreases</td>
</tr>
<tr>
<td>$\beta$</td>
<td>not monotone</td>
<td>not monotone</td>
<td>not monotone</td>
<td>increases</td>
<td>increases</td>
</tr>
<tr>
<td>$\delta_i$</td>
<td>decreases</td>
<td>decreases</td>
<td>decreases</td>
<td>decreases</td>
<td>decreases</td>
</tr>
<tr>
<td>$v_i$</td>
<td>increases</td>
<td>increases</td>
<td>decreases</td>
<td>no change</td>
<td>decreases</td>
</tr>
</tbody>
</table>

Another difference between the results for this extension and those of the base case model is that the optimal wait-time for the monopolist now depends upon the price - see Table 2.2. This is true even for the special case when $\delta = 0$. In this case, a higher price indicates a higher optimal waiting time and the out-of-pocket payment does not impact the optimal wait time (whereas in the base case neither the price nor the out-of-pocket payment impact the wait time). This result is somewhat surprising, especially since we demonstrate that lower reimbursement prices could actually lead to lower waiting times as opposed to increase moral hazard leading to higher demand and higher waiting times. We offer an explanation assuming $\beta > 0$ and $q > r\delta$, - i.e. higher quality effort improves quality and slows service speed or improves service speed at high cost. In this case, lower reimbursements make quality effort relatively more expensive, and therefore the provider chooses a lower quality effort. Since quality effort is reduced in equilibrium, more consumers choose the outside option and wait times decrease as consumers are not willing to wait as long for lower service value. While the equilibrium wait-time is lower, quality has also been effectively reduced.
Table 2.3: Summary of Impact of Competition on Quality, Access, Social Welfare

<table>
<thead>
<tr>
<th>Impact of Competition</th>
<th>On Quality Effort</th>
<th>On Wait-Time</th>
<th>On Social Welfare</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base Model</td>
<td>duplication</td>
<td>no impact</td>
<td>-</td>
</tr>
<tr>
<td>Utilization-Dependent</td>
<td>reduction</td>
<td>lower</td>
<td>+/-, -</td>
</tr>
<tr>
<td>Technical Quality</td>
<td>reduction</td>
<td>lower</td>
<td>-</td>
</tr>
<tr>
<td>Quality cost per arrival</td>
<td>escalation</td>
<td>higher</td>
<td>-</td>
</tr>
<tr>
<td>Quality cost per unit of capacity</td>
<td>escalation</td>
<td>higher</td>
<td>-</td>
</tr>
<tr>
<td>Single Decision Variable</td>
<td>reduction, duplication</td>
<td>lower, no impact</td>
<td>+, - *</td>
</tr>
</tbody>
</table>

*Duplication of quality effort will consistently correspond with social welfare loss, but reduction in quality effort due to competition can correspond to settings with either a social welfare gain or loss.

2.7 Summary

We have presented various models to show how competition under fixed prices may impact quality, access, and social welfare. Table 2.3 summarizes the key results of the six different settings for the model. In our base model we find that competition leads to duplication of quality effort and loss of social welfare which supports the duplication, but not the escalation side of the Medical Arms Race. We show other model extensions where competition leads to escalation. However, an important overall takeaway is that under our relatively simple, stylized assumptions, we show that competition can lead to lower quality effort and wait-time at a given provider whereas the previous literature consistently supports duplication or escalation of quality effort due to competition. We find that incorporating wait-times may present an opposing argument to the Medical Arms Race argument and previous economic theory, and we show how it is important to consider both the consumer’s valuation for service and the wait-time for service.
2.8 Conclusions & Future Work

In this work, we use a competitive queuing model to compare quality, access, cost, and social welfare between monopoly and duopoly equilibria with fixed prices. Under the assumptions of homogenous consumers, exponential service times and additive linear costs, we find that any entering provider(s) always captures the full market and the optimal quality and wait-time at a given provider are independent of whether a competitor enters the market. Consumers follow a queueing equilibrium strategy and their utility does not change due to competition. When two symmetric firms are in the market, a duplication of quality efforts and capacity buffers results in a corresponding loss of social welfare. The monopolist setting corresponds to the industry profit-maximizing and socially efficient outcome. These results agree with the medical arms race argument with regard to duplication, but not escalation. If policy makers wish to maximize social welfare, then competition leads to poor performance and the base model supports restricting market entry for certain services through programs such as Certificate of Need.

We find that competition may lead to either higher or lower quality at a given provider, and higher or lower social welfare under fixed-price competition when we allow for alternative model assumptions on the relationship between quality and capacity. The extensions to the model offer arguments both for and against the established result from economic theory that fixed-price competition will consistently lead to higher quality effort. We find that quality effort and wait-time will decrease due to competition in cases where higher quality effort leads to higher technical quality, i.e. higher probability of successful outcome. For example, in the inpatient care scenario, higher quality effort may imply longer patient stays but also fewer readmissions for patients who were incorrectly or incompletely treated on an initial visit. In the outpatient
scenario, higher quality effort could also mean a more complete and thorough diagnosis of the problem which would minimize potential rework in the future. In these settings we find that competition leads to lower wait and lower quality effort. Alternatively, quality effort and wait-time will increase due to competition when the cost of quality effort is volume based, either in terms of per arrival or per unit of capacity. We find that competition may decrease quality and wait-time at a given provider when the quality and capacity decision are driven by a single choice, e.g. when base capacity is fixed.

Numerous insights for managers can be found within our work. First, managers should balance the costs of quality and capacity such that when marginal cost of quality increases (above its original value), providers should respond by decreasing quality effort and increasing capacity which improves access through lower wait-time. If the cost of capacity increases, providers should respond by increasing quality effort and wait-time (decreasing capacity). This balance of the costs of quality and access connects with the idea of the “Iron Triangle” of health service delivery in that relative cost tradeoffs determine the optimal allocation between quality and access. Second, our sensitivity analysis shows that if consumer-wait time sensitivity increases, providers increase quality and capacity (lower wait-time). Therefore, as consumers become increasingly informed about wait-time expectations, providers should increase capacity to keep wait times low, but also increase both quality effort and capacity to offset the loss of service value and make waiting for service more worthwhile from the patient’s perspective.

Our research also has implications for policy makers. First, the loss of social welfare due to competition exists even when consumers are not wait-sensitive. The loss of social welfare increases as wait sensitivity increases. This is due to a higher capacity buffer at each provider and a corresponding higher level of buffer duplication. Second, to maximize social welfare or
minimize the social welfare loss due to competition, policy makers should set consumer out-of-pocket payments low. Lowering out-of-pocket payment, however, will lead to a decrease in optimal quality in both the monopoly and duopoly settings with no effect on optimal expected wait time. Higher out-of-pocket payments may decrease moral hazard, but this may not imply lower wait times, as providers respond by increasing quality effort which leads to a higher arrival rate for a given price. Third, when quality and capacity are independent decisions, the reimbursement price only impacts profitability but not quality or access. When quality and capacity decisions are correlated by a single decision, we find that higher reimbursement prices lead to higher wait times. The implication for policy makers is that reducing reimbursement prices may bring about savings in overall expenditures for payers. This could lead to reduced quality effort from the providers but with the benefit of reduced wait-time for service.

Our work offers multiple opportunities for future research. One primary avenue for future research is to consider multiplicative heterogeneity among customers, i.e. cases where patients have a heterogeneous sensitivity to quality, e.g. some patients prefer new amenities more than others or some patients are more sensitivity to side effects of treatment than others. Furthermore, future research should consider how an underlying health state may lead to heterogeneity among multiple components of the consumer utility function: quality sensitivity, wait sensitivity, service time, service value, etc. For example, patients in a more urgent or critical condition, may receive higher valuation from service (because the pain relieved and loss of productivity restored by treatment could be potentially greater) and also be more sensitive to waiting time. Future work might also examine a model where consumers may observe queue lengths or wait-time signals before making decisions regarding which queue to join. For example, patients may call to check appointment delays or backlogs with specialty providers before confirming which
specialist to schedule for service. Two ongoing works by the authors examine a provider pricing model as well as vertical competition between the provider and the third-party payer which sets the coinsurance rate to minimize its costs.
Chapter 3

Capacity Planning for Cancer Prevention

3.1 Introduction

Colorectal Cancer (CRC) is a significant health concern both in the US and worldwide. The National Cancer Institute estimates 146,970 new cases in 2009 with 106,100 of those being colon cancer (National Cancer Institute 2013b). CRC is the second leading cause of cancer death in the United States (National Cancer Institute 2013b). Colorectal cancer is a cancer which begins in the colon or rectum, typically as a small polyp. Risk factors for CRC, ranging from diet, incidence of smoking, and family history of cancer, have been identified through various empirical studies (Walsh and Terdiman 2003). Age is an important risk factor. After 50 years of age the incidence of polyps in the colon increases exponentially for the general population. Less than 5% of all cases are diagnosed before age 44 while the median age at diagnosis is 70. CRC has an asymptomatic disease progression in that the symptoms associated with colon cancer are difficult to distinguish from other more common ailments and many patients show no pain or symptoms. Diagnosis relies primarily on various screening techniques. As with other cancers, there are benefits of early detection due to the effectiveness of treatments in reducing mortality when the disease is detected at an early stage. One example is that the 5-year relative
survival rates are 90.3% for localized CRC; 70.4% for regional CRC, and 12.5% for distant CRC (National Cancer Institute 2013a).

Colonoscopy has become a standard screening procedure for colorectal cancer in people who are over 50 because it has few risks and it is highly effective (National Cancer Institute 2013c). Colonoscopy is a type of endoscopy where a flexible colonoscope is passed through the anal canal into the rectum and colon. If the test does not detect abnormal growth of polyps, future colonoscopies may be scheduled at 5 to 10 year intervals. Alternative screening options are also available which include, but are not limited to, fecal occult blood testing (FOBT) and sigmoidoscopy. Colonoscopy, because it is a more thorough exam, is more effective and more costly ($550 to $1500), but is also more likely to include greater risks and discomforts to the patient, such as sedation (Leshno et al. 2003). Sigmoidoscopy differs from colonoscopy in that colonoscopy allows an examination of the rectum and entire colon (five feet) whereas sigmoidoscopy only allows an examination of the lower colon (two feet). Sigmoidoscopy ranges from $200 to $400, and though it causes some discomfort for the patient, it is usually performed relatively quickly without sedation (Leshno et al. 2003). FOBT testing is not as effective in diagnosing CRC with an estimated sensitivity rate of 50% and a specificity rate of 90%; however, it is a widely available, quick, and inexpensive test ($5-$35) (Leshno et al. 2003). New tests such as virtual colonoscopy, or CT assisted colonoscopy, and stool based molecular testing expand the screening possibilities but are less common in practice because of costs, availability of new technology, and comparative lack of empirical validation compared to colonoscopy (Walsh and Terdiman 2003, Markowitz and Bertagnolli 2009, Link et al. 2010, Dachman and Laghi 2011).

None of the tests are 100% accurate and stakeholders including patients, providers, and insurers must balance the cost of the test against the accuracy of the test in determining which screen, if
any, to use. The National Cancer Institute provides a publicly available comprehensive summary of the methods for colorectal cancer screening and the evidence of benefit for each (National Cancer Institute 2013c). While the accuracy of a colonoscopy may vary based on the subjective judgment of the physician, it remains the most common, thorough and well-established screening exam. Colonoscopy is often used to confirm diagnosis of suspected CRC in patients who show positive symptoms or results from other screens (National Cancer Institute 2013c). Therefore, colonoscopy screening is the focus of this work.

Despite well-established guidelines, adherence to CRC screening is a public health concern in the US and worldwide. The guidelines endorsed by the Center for Disease Control and Prevention (CDC) and the US Preventive Services Task Force (USPSTF) recommend routine screenings beginning at age 50 following one of three plans: 1) FOBT every year; 2) Sigmoidoscopy every 5 years with FOBT every 3 years; 3) Colonoscopy every 10 years (CDC 2013b). The CDC guidelines describe how colonoscopy is also used as a diagnostic test when a person has symptoms or the results of another colorectal cancer screening test are unclear or abnormal. Subramanian et al. (2004) estimate that 34% of the US population obtained recommended screening in 2000. Taylor et al. (2011) find that 48.4% of those 60-69 had a colonoscopy in the last ten years. Inadomi et al. (2012) estimate 38% adherence rate for colonoscopy guidelines.

Numerous studies identify factors, including both patient and health system characteristics, which drive patient adherence to CRC guidelines and levels of CRC testing. These factors shed light on interventions which could potentially improve colonoscopy adherence. Physician recommendation for testing and patient education regarding CRC screening are found to be significant drivers across several studies (Brawarsky et al. 2004, Myers et al. 1991, Vernon 1997, Zapka et al. 2002). Patients may be informed about the need for CRC screening by their primary
care provider, local health clinic, or insurance provider. Screening information may also come from mass-media campaigns, e.g. the “Katie Couric” effect refers to an estimated 20% increase in the rate of colonoscopies across America after Katie Couric’s televised colonoscopy in 2000 (Cram et al. 2003). Studies have identified demographic and other patient-level characteristics such as age, ethnicity, family history of the disease, and type of insurance coverage as impacting a patient’s screening adherence level (Walsh et al. 2002, Brawarsky et al. 2004, Zapka et al. 2002). Some studies find fear of pain or embarrassment to be a factor which influences patient adherence to guidelines and screening recommendations (Brawarsky et al. 2004).

Within the health system, there are also financial barriers to why some do not choose to obtain routine colonoscopies. The price of the service can be a limiting factor for some, especially those without insurance coverage or those with insurance coverage which does not cover screening. A 2013 article in the NY Times, “The $2.7 Trillion Medical Bill: Colonoscopies Explain Why U.S. Leads the World in Health Expenditures” describes how colonoscopies are the most expensive screening test that healthy Americans routinely undergo and that many are unaware of the overall amount billed or their out-of-pocket cost until after scheduling the procedure (Rosenthal 2013). Figure 3.1 from the article shows how the cost of a colonoscopy varies by region and cites the use of an anesthesiologist as one explanation for the variation in cost (Rosenthal 2013). Liu et al. (2012) find an increase in the the level of colonoscopies over time as well as increase in the use of anesthesia services with colonoscopy. They find that from 2003-2009, gastroenterology procedures per million enrollees remained relatively stable in the Medicare population but increased more than 50% in commercially insured patients. In both populations the use of anesthesia services increased from approximately 14% to 30% over the same time frame, and there was substantial regional variation in the proportion of proce-
tures using anesthesia (13% to 59%). Payments for anesthesia services doubled in Medicare patients and quadrupled in commercially insured patients. Better understanding the relationship between various factors, including patient population characteristics, disease progression parameters, and economic drivers, and how these factors together impact operational decisions such as capacity, may help to explain regional variation with respect to cost and access.

Another barrier to colonoscopy adherence is appointment access. In 2002 and 2003, two NY Times articles chronicled this problem with quotes and anecdotes from various health professionals facing the challenges surrounding long delays for colonoscopies (Scott 2002, Kolata 2003)

“Healthy patients at the center cannot have colonoscopies because the waiting lists are closed...It’s fine to say everyone should have a colonoscopy,” Dr. (John H.) Bond (VAMC Minneapolis) said. “But we are talking about 70 million people. It is
unclear whether that is even feasible in the United States.” - (Kolata 2003)

“If you’re urging people to be screened and then you say, O.K., the colonoscopy will be a year from now, you shoot yourself in the foot,” said Dr. Robert H. Fletcher, (Harvard Med.). “The meta-message from the health care community is, well, it’s not that important after all.” - (Kolata 2003)

In the year since Medicare began paying for wider access to colonoscopy to look for colon cancer (2001), the number of people having the test has greatly increased, and doctors say they are struggling to keep up with demand. Dr. Michael Pignone, (UNC-CH), said the average waiting time for an appointment for a colonoscopy was three to six months. - (Scott 2002)

In DeForest, Wis., 20 minutes from Madison, the waiting lists for the colonoscopies are closed to healthy patients. Dr. Peter Pickhardt, a family practitioner, says he has learned to be blunt with patients. “I tell them up front,” he said. “If they want a colonoscopy, it’s not available.” - (Kolata 2003)

While increased education about the guidelines is a key to adherence, it is unclear whether current system capacity could even sustain the demand generated if all patients were to follow the guidelines for routine screening. The concerns around a lack of supply are magnified by shifts in the population distribution which lead to more patients over age 50, and also major health policy reform which requires insurers to cover certain preventive services, including colonoscopy screening.

Little is understood about how screening guidelines and capacity planning decisions are related in practice. The question of how often an individual should be screened for various forms of cancer has been well-documented in the literature from the perspective of optimizing
an individual’s quality-of-life tradeoffs or comparative effectiveness analysis (Alagoz et al. 2011, Walsh and Terdiman 2003). However, service system design and control issues surrounding cancer screening have received less research attention. When considering a market or system-level approach to this problem, it is important to consider that an individual patient’s screening behavior impacts not only the cost and disutility for that patient, but the access to service, system load, and wait time for service experienced by other patients requesting screens. As a contribution to the established medical literature on CRC screening, our model investigates system-level dynamics which clinical screening optimization models for individuals may ignore.

Patient adherence to colonoscopy screening is not simply split into those who adhere to the guidelines and those who do not, but can be more accurately described as a continuum of how well patients adhere to the guidelines (i.e. how much longer do they wait between screens than is recommended). Some patients may wait so long that they die from other causes before ever being screened. Other patients may request a screen immediately after it is recommended. We capture the range of imperfect patient adherence behavior by modeling the time patients request the next screen to be a random variable. This heterogeneity with respect to the time patients spend waiting to schedule the next screen represents variation in patient adherence to the guidelines which could be driven by variation in patient education and information throughout the population.

In this work we develop a queuing network for colorectal cancer screening in order to 1) analyze how a social planner should optimize cancer screening capacity in order to minimize the sum of detection delay and capacity operating costs; 2) examine what tradeoffs exist between the capacity planning decision and the time between screens (i.e. screening guidelines). In Section 2, we provide a review of the relevant literature spanning from empirical work on the
comparative effectiveness of CRC screening to operations research models for cancer screening. In Section 3, we describe our model and provide some preliminary analysis to characterize the model dynamics. In Section 4, we address the first research problem, by studying the capacity planning decision when the average time until patients schedule the next screen is constant, i.e. screening guidelines and patient adherence are relatively stable. We address the second research problem in Section 5 and Section 6. In Section 5, we assume that the screening capacity is constrained, but the social planner may determine the average time until patients schedule the next screen. We examine sensitivity of the optimal time until the next screening request to changes in capacity. In Section 6, we investigate a model where the social planner minimizes total cost by simultaneously optimizing screening capacity and the average time until patients schedule a follow-up screen. The joint optimization allows further investigation of the tradeoffs between the two decisions. In Section 7, we perform a numerical study using parameter estimates derived from public health data in order to 1) relax special case assumptions needed for some parts of earlier analysis; 2) estimate capacity, average time until scheduling the next screen, total cost, and wait time based on public health data; and 3) quantify the effects of parameters on optimal decisions and system performance. In Section 8, we present conclusions of our research and describe opportunities for future research.

3.2 Literature Review

Our work relates to literature on empirical studies which assesses the impact, value and cost effectiveness of colorectal cancer screening and analytical models for health screening.

An extensive body of research covers empirical study of the value and cost effectiveness of colorectal cancer screenings. Walsh and Terdiman (2003) provide an extensive review of
evidence-based literature on colorectal cancer screenings. Sonnenberg et al. (2000) develop computer models of a Markov process to evaluate three screening policies, annual FOBT, flexible sigmoidoscopy, and colonoscopy. The authors find that annual FOBT results in higher cost savings but saves fewer life-years than colonoscopy. The authors also use sensitivity analysis to investigate the impact of patient compliance, finding that colonoscopy is the most cost-effective primary screening strategy for low compliance. Frazier et al. (2000) also use a Markov model to compare 22 different CRC screening strategies, 7 non-dominated, under the outcomes of life expectancy and discounted lifetime costs. The authors allow for variation in the compliance rate and find that even with imperfect compliance, CRC screening significantly reduces mortality at costs comparable to other cancer screenings. Of the policies considered the authors find that FOBT plus sigmoidoscopy every five years is the most cost effective strategy.

There is also a significant body of literature on the application of mathematical decision models to the health screening decision. Kirch and Klein (1974) develop an early model for determining screening schedules in age-dependent diseases such as colorectal cancer. The authors seek to choose the optimal number of examinations over a fixed time period (frequency) in order to minimize the (expected) detection delay. A key assumption of their models is that examinations are performed at equal intervals. Another limitation is that they assume that examinations are 100% effective. Despite these limitations, the authors formulate a simple model for which they can derive the optimal schedule explicitly as non-periodic. This means that the optimal number of examinations is proportional to the square root of the age-specific incidence probability of the disease.

Stochastic dynamic programming is a common method of analysis used to model the health maintenance decision in the literature. Alagoz et al. (2011) review operations research models
for cancer screening examining 41 papers and 33 models. Ozekici and Pliska (1991) develop a semi-Markov model for the optimal scheduling of health inspections with false positives and negatives. The authors use a cost minimization framework with total costs being comprised of inspection costs, false-positive costs, false-negative costs, and true-positive costs associated with onset of the disease. A limitation of the model is that the costs of a false-positive or a false-negative have less tangible meaning than other reward functions such as quality-adjusted life years or mortality. Therefore, these costs can be difficult to estimate in practice. Hauskrecht and Fraser (2000) use a Partially-Observable Markov Decision Process (POMDP) to formulate the problem of treating ischemic heart disease. The characterization of hidden disease states allows them to investigate treatment procedures and cost-benefit trade offs for different policies. Maillart et al. (2008) use a POMDP to investigate various mammography screening policies. The authors formulate an efficient frontier of policies with respect to lifetime mortality and mammography count. The authors use a policy evaluation approach, as opposed to optimal policy evaluation, so that they may compare current policies in practice and provide a menu of efficient policies from which an individual patient / decision maker may choose a preferred option. Zhang et al. (2012) develop a POMDP for prostate cancer screening with two available actions: treat (based on screening results) or wait. The authors extend previous research by establishing theoretical results such as conditions for the existence of a control-limit policy. A limitation of this model is that it does not integrate the screening decision and the treatment decision. In general these models view the screening decision from the perspective of a single patient without taking into consideration the screening behavior of other patients. Also, the authors do not study the effect of imperfect patient compliance with the suggested guidelines.

Many operations research models have employed simulation or stochastic dynamic program-
Clemen and Lacke (2001) perform decision analysis using a deterministic growth model for colorectal cancer and Monte Carlo simulation to address uncertainty about model parameters. The authors also perform utility analysis to find that colonoscopy every three years is the top-ranked strategy of those they considered. These results differ from the empirical findings above which typically refer to colonoscopy every 10 years or FOBT and sigmoidoscopy every 5-10 years. Leshno et al. (2003) use a POMDP to study colorectal cancer screening policies. The authors perform a detailed cost-effectiveness analysis of several standard policies using the POMDP; however, they do not address the optimal policy or heuristic policies which may be an improvement over current practice. Another limitation, is that there is no optimization of the time between screenings; for instance, the authors assume 10-year intervals for colonoscopy as the given guideline. The model could be also extended to integrate multiple treatment decisions and objectives.

The primary contribution of this research will be to examine how social planners should plan capacity for a screening population as opposed to using a medical decision making model to optimize the screening decision for an individual patient within the population. An important, relevant distinction is that in our model patient wait-times (appointment delays) impact the time until patients can access screening and the time until those with cancer can be diagnosed and begin treatment. These delays will result in higher treatment costs, lower quality of life, and increased mortality for those in whom cancer is detected. The cost to society of screening an individual does not depend solely on the monetary cost to that individual but also the increased system load and subsequent waiting cost incurred by other patients. We formulate an objective function for a social planner which minimizes the cost tradeoffs between the detection delay
costs and capacity costs associated with a cancer screening facility. We use public health data for CRC screening to calibrate our model and inform government planning and intervention decisions such as how much to invest in additional endoscopy suite capacity or what efforts should be made to influence the average time until patients schedule the next screen. Our model is not meant to supersede other models developed in the clinical and medical decision making literature which are used as a basis for the screening guidelines, because it employs more stylized assumptions. However, our model can provide support for those guidelines or reason to investigate them further. It can also be used to judge the economic implications of how much capacity is needed to ensure adherence to established guidelines can be accommodated with relatively low wait times.

3.3 Model Description

In this section we present a model for optimizing the capacity decision for cancer screening. Figure 3.2 illustrates the model setting. We will describe our model as a queueing network with 2 nodes: 1) Home; 2) Endoscopy suite. Patients cycle between these two nodes until either they die from other causes or cancer is detected via screening or other methods. We will let the total input rate into Node $i = 1, 2$ (number of patients arriving to a node per unit time) be given by $a_i$. New patients enter the system once they reach the minimum age for routine screening as a Poisson process with rate $\lambda$ - e.g. $\lambda$ represents the number of persons within the endoscopy suite’s market who turn age 50 in a given week. Each patient enters with three random variables which correspond to the patients’ health states. Let the random variable $X_n$ denote for patient $n = 1, 2, \ldots$ the time after entry into the system until cancer is detectable via screening. For all intents and purposes, we can think of $X_n$ as the time of onset of the disease.
For example, if Patient $j$ develops cancer at age 62 after entering the system at age 50, then $X_j = 12$ years. Let the random variable $L_n$ denote the time after entry into the system until patient $n$ dies (would die) from other causes. For example, if Patient $j$ dies at age 89 after entering the system at age 50, then $L_j = 39$ years. Let $Y_n$ denote for patient $n$ the time after cancer becomes detectable via screening until cancer is detected via symptoms or diagnosis by methods other than the primary screen (e.g., FOBT testing as opposed to colonoscopy). For example, if Patient $j$ develops cancer at age 62 and then finds the cancer via FOBT testing at age 66, then $Y_j = 4$ years. The inherent assumption behind $X_n$ and $Y_n$ being continuous random variables for all patients is that if patients lived forever, eventually everyone would get colorectal cancer and eventually it would be found via other methods. However, the incidence rate of colorectal cancer is relatively low compared to the mortality from other causes, so it is likely that in relevant applications of the model $X_n$ will be much larger than $L_n$ for most patients. Also, some patients will call to request a screen before, $X_n + Y_n$, the time when cancer is found via other methods, and the cancer will be found via screening first.

Given these three health states we now describe the time line for an entering new patient. The patient $n = 1, 2, ...$ spends a random amount of time at the Home node, $H_{1n}$ before attempting to schedule her first screen. During the time at home, the patient may die from other causes if $L_n = \min (H_{1n}, L_n, X_n + Y_n)$ or detect cancer via other methods if $X_n + Y_n = \min (H_{1n}, L_n, X_n + Y_n)$. If either event happens, the patient leaves the system. If the patient has not died from other causes or detected cancer via other methods by the end of the time at home, she schedules the first available appointment and joins the queue for the endoscopy suite where the total colonoscopy throughput time (sum of appointment delay and service time for colonoscopy) is given by $W_{1n}$. After service, either cancer is detected and the patient leaves the
system to begin treatment, or cancer is not detected and the patient returns to the home stage where the cycle repeats. The total arrival rate into the home node, $a_1$, is equal to the sum of the new patient arrival rate and the rate of patients who were screened for cancer with negative results and return to the home stage. For subsequent cycles, we generalize the above notation as follows: let $H_{in}$ denote the time until patient $n$ schedules screen $i$; $W_{in}$ denote the appointment delay for screen $i$ for patient $n$; $S_{in} = H_{in} + W_{in}$ denote the total cycle time between screen $i-1$ and screen $i$ for patient $n$. (e.g. $H_{2n}$ denotes the time patient $n$ waits at “home” before scheduling her second colonoscopy screening appointment).

We make the following distribution assumptions on the above random variables. We model the home node as having an infinite number of servers each with an exponential service-time distribution with mean $\frac{1}{\tau}$. The infinite number of servers assumption implies that patients do not wait to enter the “home” stage, but simply spend a random “service time” there until
requesting the next screen - i.e. patients have “self-service” at home. Therefore, \( H_{in} \) follows a common exponential distribution with mean \( \frac{1}{\tau} \). We model the endoscopy suite as a single server with service rate \( \mu \). The single server is an approximation of the aggregate clinic process. The variability in service time refers to aggregate variability as opposed to variability at any given stage within the clinic. Given this assumption, \( W_{in} \) follows an exponential distribution with mean \( \frac{1}{\omega} = \frac{1}{\mu - a_2} \); recall that \( a_2 \) is the total input rate (demand per week) for screens. We assume that the time until cancer is detected in patient \( n, X_n \), is exponentially distributed with mean \( \frac{1}{\beta} \). This means that given a patient has not developed cancer by age \( t \) the probability of developing within the next \( \delta \) time units is the same regardless of the time expired, \( t \). We assume that the lifetime random variable, \( L_n \), is exponentially distributed with mean \( \frac{1}{\alpha} \). We assume that the time until cancer becomes detectable via other methods once detectable via screening, \( Y_n \), is exponentially distributed with mean \( \frac{1}{\theta} \). We assume that these distributions are common to the entire patient population. Assuming the exponential distribution of the patient health random variables implies age-independence with regard to mortality rate and cancer rate. We discuss below the limitations of and justification for the exponential distribution as well as the assumption of a common distribution for all patients.

There are two sets of Markov chains driving the system dynamics. The first Markov chain describes the queueing network and patient flow between screening and waiting to be screened. This is characterized by a birth and death process at each of the two nodes. The second set of Markov chains describes the health state of the patient who at any time can be healthy, have cancer undetected, have cancer detected via screening (and left the system), have cancer detected via other methods (and left the system), or have died from other causes (and left the system). The chain is a continuous Markov chain, i.e. the patient spends a random time in each
of the health states and the random variables, \( X_n, L_n, \) and \( Y_n \) and the events specified below describe the transitions between the possible states for the patient. Under general assumptions on the random variables, a complete characterization of the state of the system requires tracking every patient in the system, each patient’s health state, and whether they are at the home node, in service at the endoscopy suite, or waiting to be screened. The exponential distributions allow for great simplification of the state space. Due to the memoryless property of the distribution, each patient’s cycle time between screens (time at home plus time waiting for screen) and health random variables will regenerate upon any entry into the home node. Therefore, we can drop the subscripts on the random variables as any given screening cycle will have the same distributions across patient and cycle number.

Now we can describe the dynamics of the system using four possible events which may happen to any patient in any given cycle: 1) death from other causes while at home \((DH)\); 2) other detection of cancer while at home \((SH)\); 3) screen which tests positive for cancer developed at home \((CH)\); 4) screen which tests negative for cancer \((NC)\). For analytical tractability we assume that the wait time for an appointment is relatively small compared to the time spent at home (e.g. 10 weeks vs. 10 years) such that the probability of a patient dying from other causes, detecting late stage cancer via other methods, or developing cancer while waiting to be screened is negligible. In order to make this assumption we reassign some of the possible sequences of the random variables and assume the process regenerates upon completion of the screen. For a given cycle, if a patient dies at home, i.e. event \(DH\) occurs, we assume she leaves the system without cost to the social planner. If cancer is detected via other methods within a cycle, i.e. event \(SH\) occurs, a patient also leaves the system but the social planner incurs a penalty cost \( b_2 \) per unit of time during the detection delay (the time from the onset of cancer, \( X_n \), to the time
$X_n + Y_n$ when cancer is detected via other methods). If the patient does not die or have cancer detected via other methods while at home, she calls to make a screening appointment and will wait to be screened. If the screen detects cancer, i.e. event $CH$ occurs, then the patient leaves the system and the social planner incurs a penalty cost $b_1$ per unit of time during the detection delay. If the screen does not detect cancer, i.e. event $NC$ occurs, then the patient returns to the home node and the cycle regenerates due to the memoryless distribution.

The social planner wishes to minimize the total expected cost per unit time, $TC$. As above, let $b_1$ denote the detection delay cost per unit time for patients whose cancer is found via screening and $b_2$ denote the detection delay cost per unit time for patients whose cancer is found via other methods. These parameters represent how the cost of treatment, disutility from treatment, and risk of mortality from cancer are increasing in the time between the onset of the disease and the time the disease is detected. We assume that these delay costs are linear, whereas in reality, it could be that the delay costs increase at an increasing rate - e.g. delaying a patient’s treatment from 4 to 8 weeks might have a much smaller cost than delay from 64 to 68 weeks when the cancer is more likely to be spreading to a distant stage more rapidly. Though not a restrictive assumption of the model, it is likely for our application of the model that $b_1 \leq b_2$ because the cancer grows relatively asymptotically. If CRC is detected via other methods, it is likely detected in a much later stage with high treatment costs. Let $c$ denote the marginal cost of capacity per unit time; we assume a linear cost of capacity for simplicity as the results will not differ dramatically if we use a convex cost function (e.g. quadratic). The marginal cost of capacity, $c$, represents the average cost per time unit (e.g. week) of increasing the capacity by one. We write the long-run average total cost function, $TC$, as follows
The motivation for using this model is to gain insights into the operational issues regarding capacity planning and guideline decisions for colorectal cancer screening. The model has several limitations as it is a stylized representation of reality. One important limitation of the model is that we do not explicitly model the treatment of patients once cancer has been detected. We also assume that patients who have been treated for cancer do not return to the screening process. In reality, patients who have been treated for CRC will continue to be screened, often times even more frequently than those with no previous diagnosis. This would require clinics to carry additional screening capacity than is estimated by our model.

A second limitation of the model is that we make extensive use of the exponential distribution for our assumptions regarding randomness in the model, including: time until death, time until cancer is detectable, time until cancer is discovered via other methods, time in queue, and time at home. In reality, the exponential distribution may over-estimate the variability for some of the randomness described by the model. In terms of estimation, the exponential distribution only allows one parameter for scale and not one for shape. The remaining lifetime for the exponential distribution is time-independent whereas in reality, rates such as the mortality rate may increase as patients age. These changes in mortality rate due to age of the patient are less of a concern in a long-run planning model compared to a short-range scheduling or capacity allocation decision model. A multi-class model would be a natural extension of our work which could incorporate such heterogeneity in patient characteristics.

Another limitation is that we look at the model from the perspective of a long-term steady-
state, whereas in practice, regional population growth or shifts in the population age distribution may cause changes in the arrival rate over time (e.g. current aging of the baby boomers). Since our model is a high-level strategic planning model, we are able to obtain relevant insights regarding changes in the arrival rate from our sensitivity analysis. Another limitation of the model is that we assume that a screen for cancer is completely accurate in terms of 100% sensitivity and specificity. In reality, whether a colonoscopy detects colorectal cancer within a patient is subject to the proficiency of the provider and the test may not be entirely accurate (National Cancer Institute 2013c). This assumption is still appropriate for the level of decision-making considered in the model and the reality that colonoscopy is by far the most thorough and accurate form of colorectal cancer screening available (National Cancer Institute 2013c). A final limitation is that the model does not account for heterogeneity among patients with respect to risk factors for the disease including gender, ethnicity, and family history.

### 3.3.1 Preliminary Analysis

Using the memoryless property of the exponential distribution and our simplifying assumptions, we derive Lemma 3.1 which characterizes the probability of the four events within any given cycle.

**Lemma 3.1.** The probability of the four events within a cycle are given by 

\[
P(DH) = \frac{\alpha(\tau+\beta+\alpha+\theta)}{(\tau+\alpha+\beta)(\tau+\alpha+\theta)};
\]

\[
P(SH) = \frac{\beta\theta}{(\tau+\alpha+\beta)(\tau+\alpha+\theta)};\quad P(CH) = \frac{\beta\tau}{(\tau+\alpha+\beta)(\tau+\alpha+\theta)};\quad P(NC) = \frac{\tau}{\tau+\alpha+\beta}
\]

**Proof.** All proofs in the Appendix.

Note that our simplifying assumption implies that the transition probabilities will not depend on the mean wait time (though the delay costs will depend on the mean wait time). Using the
above probabilities we write the balance equations for the described Jackson Network as follows.

\[ a_1 = \lambda + a_2 P(\text{NC} | \text{Screen}) \]  
\[ a_2 = a_1 P(\text{Screen}) \]

Substituting the transition probabilities using Lemma 3.1, we solve the balance equations and the results are given in Lemma 3.2.

**Lemma 3.2.** Solving the balance equations, 3.2.3.3, we obtain the following expressions for the steady-state input rates of the queueing network: 

\[ a_1 = \frac{\lambda(\tau + \alpha + \beta)}{\alpha + \beta}; \quad a_2 = \frac{\lambda(\tau + \beta + \alpha + \theta)}{(\alpha + \beta)(\tau + \alpha + \theta)}. \]

In order to calculate the long-run average total cost function we need to calculate the expected delay when cancer is detected via other methods and the expected delay when cancer is detected via screening. Conditioning on all possible sequences of the random variables within a cycle we derive the expected delay cost per cycle, \( DC \), as follows. We let \( \frac{1}{\omega} = \frac{1}{\mu - a_2} \) denote the mean wait time.

\[ DC = \frac{b_1 a_1 \tau \beta (\alpha + \tau + \theta + \omega)}{\omega (\tau + \alpha + \theta)^2 (\tau + \alpha + \beta)} + \frac{b_2 a_1 \beta \theta}{(\tau + \alpha + \beta)(\tau + \alpha + \theta)^2} \]  
\[ (3.4) \]

Evaluating the above expression at \( \omega = \mu - a_2 \) and substituting for \( a_2 \), we write the long run average delay cost per unit time, \( DC \), as follows:
\[ DC = \lambda \left( \frac{\beta}{\alpha + \beta} \right) \left( \frac{\tau}{\tau + \alpha + \theta} \right) \left( \frac{b_1}{\tau + \alpha + \theta} + \frac{b_1}{\mu - \frac{\lambda \tau}{(\alpha + \beta)(\tau + \alpha + \theta)}} \right) \]
\[ + \lambda \left( \frac{\beta}{\alpha + \beta} \right) \left( \frac{\theta}{\tau + \alpha + \theta} \right) \left( \frac{b_2}{\tau + \alpha + \theta} \right) \] (3.5)

Assuming a linear cost of capacity per unit time we write the provider’s optimization problem.

\[
\min_{\mu > a_2} TC = \lambda \left( \frac{\beta}{\alpha + \beta} \right) \left( \frac{\tau}{\tau + \alpha + \theta} \right) \left( \frac{b_1}{\tau + \alpha + \theta} + \frac{b_1}{\mu - \frac{\lambda \tau}{(\alpha + \beta)(\tau + \alpha + \theta)}} \right) \]
\[ + \lambda \left( \frac{\beta}{\alpha + \beta} \right) \left( \frac{\theta}{\tau + \alpha + \theta} \right) \left( \frac{b_2}{\tau + \alpha + \theta} \right) + c\mu \] (3.6)

The terms in the total cost function can be explained as follows. On average, \( \lambda \) new patients arrive to the system every period. The probability that a new patient develops cancer before she dies of other causes is given by \( \left( \frac{\beta}{\alpha + \beta} \right) \). In the cycle when the patient develops cancer (at home), if it is detected via screening, the delay cost is \( b_1 \) per unit time from onset until the patient schedules the next screen and also \( b_1 \) per unit time waiting for the screen once requested.

In order for this to happen, the next event after developing cancer must be leaving the home stage, with the associated probability \( \left( \frac{\tau}{\tau + \alpha + \theta} \right) \). The expected time in the home stage until the patient requests the appointment is \( \left( \frac{1}{\tau + \alpha + \theta} \right) \). The expected time waiting for a screen is \( \left( \frac{1}{\mu - a_2} \right) \). In the cycle when the patient develops cancer, if it is detected via other methods, the delay cost is \( b_2 \) per unit time from onset until detection via other methods. In order for this to happen, the next event after developing cancer must be detection via other methods, and
the associated probability is \( \left( \frac{\theta}{r+\alpha+\beta} \right) \). The expected time in the home stage until the patient detects cancer via other methods is \( \left( \frac{1}{r+\alpha+\beta} \right) \). We assume that the provider may increase its capacity without limit by acquiring additional resources, addressing bottlenecks, or expanding the facility. The marginal cost of doing so is assumed to be linear for mathematical convenience. This assumption simplifies the model for the sake of obtaining relevant insights.

3.4 Model Analysis: Capacity Decision

We now consider a scenario where a social planner must determine the optimal capacity which minimizes the long-run average total cost (the sum of the detection delay cost and the capacity operating cost) as given in Equation (3.6). For example, a state policy maker may have a budget for cancer prevention and be faced with the decision of how much colorectal cancer screening capacity is needed at an endoscopy suite serving a particular region of the state.

3.4.1 Cost-Minimizing Capacity

We will first analyze the above model for the case where the social planner makes only the capacity decision and has no influence over the time patients spend in the home stage. This analysis applies to cases where the guidelines for cancer screening are well-established and consumer behavior is relatively insensitive to education efforts, price changes or subsidies enacted by the social planner. To proceed with the analysis, first we show that the above long-run average total cost function is convex in the capacity decision.

**Proposition 3.1.** The long-run average total cost given in (3.6) is convex in the capacity decision, \( \mu \).
Using the objective function property shown in Proposition 3.1, we derive the equilibrium capacity decision in Theorem 3.1.

**Theorem 3.1.** The long-run average total cost minimizing capacity decision is given by

$$
\mu^* = \frac{\lambda \tau (\theta + \beta + \alpha + \tau)}{(\alpha + \beta)(\tau + \alpha + \theta)} + \sqrt{\frac{\lambda \beta \tau b_1}{c(\alpha + \beta)(\tau + \alpha + \theta)}};
$$

$$
W^* = \sqrt{\frac{\beta \lambda \tau b_1}{\lambda \beta \tau b_1}}; \quad TC^* = \frac{c \lambda \tau}{\alpha + \beta} + \frac{\lambda \beta (\theta b_2 + c \tau (\tau + \alpha + \theta) + \tau b_1)}{(\alpha + \beta)(\tau + \alpha + \theta)^2} + 2 \sqrt{\frac{\beta \lambda \tau cb_1}{(\tau + \alpha + \theta)(\alpha + \beta)}}.
$$

We rewrite the optimal total cost function so that the terms in the cost function reflect expressions which are more intuitive to the model dynamics.

$$
TC^* = \frac{c \lambda \tau}{(\tau + \alpha + \theta)(\alpha + \beta)} + \lambda \left( \frac{\beta}{\alpha + \beta} \right) \left( \frac{\tau}{\tau + \alpha + \theta} \right) \left( \frac{b_1}{\tau + \alpha + \theta} \right) + \lambda \left( \frac{\beta}{\alpha + \beta} \right) \left( \frac{\theta}{\tau + \alpha + \theta} \right) \left( \frac{b_2}{\tau + \alpha + \theta} \right) + 2 \sqrt{\frac{\beta \lambda \tau cb_1}{(\tau + \alpha + \theta)(\alpha + \beta)}}.
$$

(3.7)

The first term in the total cost function is the marginal cost of capacity multiplied by the input rate, $a_2$, into screening. This cost corresponds to the capacity which must be carried to meet the demand exactly. The second term corresponds to the delay at home for patients who develop cancer (at home) that is detected via screening during the cycle the cancer is detected. The third term corresponds to the delay at home for patients who develop cancer (at home) which is detected via other methods during the cycle the cancer is detected. The capacity decision does not impact this delay cost as carrying more capacity does not influence patients to schedule appointments sooner (or at a different frequency), it simply impacts the delay for the appointment. The last term is the long-run average cost of carrying a capacity buffer to decrease wait-times which explicitly benefits those who have undetected cancer. When the marginal cost of capacity is higher, the buffer decreases, but the cost of the buffer still increases due to the
higher marginal cost of capacity. When the delay cost is higher, the buffer increases as does the cost of the capacity buffer.

Other performance measures can be derived using the above equilibrium result. One performance measure given below which is not previously defined, is the number of screens per new-patient arrival, \( N = \frac{a_2}{\lambda} \). This gives a measure of the average number of colonoscopies in a patient’s lifetime (assuming no colonoscopies before entering the system, e.g. at age 50).

\[
a_1 = \frac{\lambda (\tau + \alpha + \beta)}{\alpha + \beta}
\]

\[
a_2 = \frac{\lambda \tau (\tau + \beta + \alpha + \theta)}{(\tau + \alpha + \theta) (\alpha + \beta)}
\]

\[
N = \frac{a_2}{\lambda} = \frac{\tau (\tau + \beta + \alpha + \theta)}{(\tau + \alpha + \theta) (\alpha + \beta)}
\]

\[
P(\text{Other.Detection}) = \frac{\beta \theta}{(\tau + \alpha + \theta) (\alpha + \beta)}
\]

\[
P(\text{Screen.Detection}) = \frac{\beta \tau}{(\tau + \alpha + \theta) (\alpha + \beta)}
\]

\[
P(\text{Negative.Screen}) = \frac{\tau}{\tau + \alpha + \beta}
\]
Table 3.1: Sensitivity of Optimal Solution

<table>
<thead>
<tr>
<th>Parameter</th>
<th>$\mu^*$</th>
<th>$W^*$</th>
<th>$TC^*$</th>
<th>$N^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$: mortality rate</td>
<td>decreases</td>
<td>increases</td>
<td>decreases</td>
<td>decreases</td>
</tr>
<tr>
<td>$\beta$: cancer rate</td>
<td>not monotone</td>
<td>decreases</td>
<td>not monotone</td>
<td>decreases</td>
</tr>
<tr>
<td>$\theta$: inverse of mean time to other detection</td>
<td>decreases</td>
<td>increases</td>
<td>not monotone</td>
<td>decreases</td>
</tr>
<tr>
<td>$b_1$: delay cost for screening</td>
<td>increases</td>
<td>decreases</td>
<td>increases</td>
<td>no effect</td>
</tr>
<tr>
<td>$b_2$: detection delay cost for other methods</td>
<td>no effect</td>
<td>no effect</td>
<td>increases</td>
<td>no effect</td>
</tr>
<tr>
<td>$c$: marginal cost of capacity</td>
<td>decreases</td>
<td>increases</td>
<td>increases</td>
<td>no effect</td>
</tr>
<tr>
<td>$\lambda$: new patient arrival rate</td>
<td>increases</td>
<td>decreases</td>
<td>increases</td>
<td>no effect</td>
</tr>
<tr>
<td>$\tau$: inverse of mean time at home</td>
<td>increases</td>
<td>decreases</td>
<td>not monotone</td>
<td>increases</td>
</tr>
</tbody>
</table>

$$P(\text{Death}) = \frac{\alpha(\tau + \beta + \alpha + \theta)}{(\tau + \alpha + \theta)(\alpha + \beta)}$$  \hspace{1cm} (3.14)

3.4.2 Sensitivity Analysis

The comparative statics for the optimal capacity, wait time, total cost, and number of screens per new patient arrival are summarized in Table 3.1.

From Table 3.1 we obtain several important insights regarding the capacity planning decision. We first look at the impact of an increase in the new patient arrival rate, $\lambda$, in order to consider the impact of an aging population, and growing demand for health services due to health policy reform and increased health insurance coverage. As more new patients enter the system, i.e. a greater number of people reach the minimum screening age, the optimal screening capacity increases as we would expect. Since we are examining a cost minimization model, more new patients mean a greater total cost as they adds greater load to the network. However, as the new
patient arrivals increase, the wait-time for screening decreases. This result is counter-intuitive, especially if one considers capacity to be relatively fixed or linearly increasing in new patient arrivals. In this long-run model, we assume the provider is not capacity constrained and can set its capacity to any level. In Theorem 3.1 we see that the optimal capacity decision, \( \mu^* \), is a function of \( a_2 \), the input rate, and a capacity square-root term which is a function of the new patient arrival rate. The input rate, \( a_2 \), is increasing in the arrival rate, but the capacity buffer is also increasing in the arrival rate in a non-linear fashion. Therefore, as more new patients enter the system, the clinic chooses an even higher capacity buffer. Since the capacity buffer is increasing in the new patient arrival rate, the wait time decreases (total throughput time as well as time in queue). We can think of this result as an economies of scale benefit in terms of service delivery (better access to service). This is an interesting result given the current changes in the market for preventive health services. The result claims that as the demand grows, delays for appointments for screens should go down assuming providers can adjust their capacity accordingly. In a later section we consider what happens in cases where capacity is constrained.

Secondly, we examine the impact of changing the average time patients spend at home, \( \frac{1}{\tau} \). Changes in the mean time at home may be due to changes in the guidelines or increased education efforts by providers, policy makers, and insurers to encourage guideline adherence. When the average time spent at home between screens decreases, the needed capacity increases as patients now return to be screened more frequently. The input into the screening node increases due to more frequent screening, as does the number of screens per new arrival. Again, we see that wait times drop. This result is counter-intuitive as one might suspect that since patients are spending less time waiting at home, the clinic can afford for them to wait longer
for their appointment and still achieve the same balance of delay and capacity costs. However, we see that as the average time spent at home decreases so does the capacity buffer because the volume effect dominates. The effect of changing the average time at home on the total cost is not monotone. Let’s consider the special case of $\alpha = 0, \theta = 0$; i.e. patients do not die of other causes nor is cancer detected via other methods. In this case, the change in total cost with respect to the inverse of the mean time at home is given by $\frac{\partial TC^*}{\partial \tau^*} = \frac{c\lambda}{\beta} - \frac{b\lambda}{\tau^2}$. When patients are spending a relatively long amount of time at home, then lowering the time spent at home will decrease the optimal total cost because the risk of extremely long cancer detection delays is reduced. If patients are already being screened very frequently (the mean time at home is extremely small), then further decreasing the time spent will have little impact on the delay cost, but require significantly more capacity from the clinic and increase long-run average total costs.

Our results also provide insights regarding changes in the cancer incidence rate, $\beta$. Such changes may be due to differences in patient populations or changes in patient behavior which contribute to risk factors for the disease such as diet or environment. As the cancer incidence rate increases, the average wait time decreases. This relationship is consistently monotone though the effect of $\beta$ on the capacity decision is not. To reach an intuitive explanation, consider the special case where patients do not die from other causes. In this case, the cancer incidence rate is extremely (infinitely) high compared to the mortality rate, $\alpha$. It is straightforward to show in this case that the capacity is monotonically decreasing in $\beta$. As patients develop cancer faster, they are also leaving the system at a faster rate. In the special case, where there is no mortality from other causes and no detection from other methods, patients only leave the system through a positive screen for cancer. The increased rate of patients exiting the system dominates the effect
of a longer detection delay for those who now develop cancer sooner. This result is reinforced by another result that in the special case of $\alpha = 0$, the total cost is monotonically decreasing in $\beta$. Patients are not dying from other causes in the special case, so when the mean time until patients develop cancer decreases, patients leave the system faster. Less capacity is needed, and less cost is incurred.

When the mortality rate, $\alpha$, increases, more patients die from other causes which decreases the capacity needed for screening since patients are leaving the system faster. Wait times increase because the clinic loses economies of scale benefits due to fewer patients in the system in steady state. The total cost decreases because patients stay in the system less time on average due to increased deaths from other causes. The average number of screens per new patient arrival also drops.

As the average time until cancer is detected from other causes, $\frac{1}{\beta}$, decreases, the optimal screening capacity decreases. This result provides particular insight for colonoscopy screening where increasing the accuracy or adherence to other relatively inexpensive methods of detection such as FOBT could decrease the need for colonoscopy capacity. However, remember that we have only modeled the screening capacity of colonoscopy and not the diagnostic capacity. This result may not hold in cases where a colonoscopy is still performed in order to confirm the presence of cancer when suspected via other methods. Also, we are not taking into account how the cost of other detection methods may change with improved accuracy. For example, molecular based stool testing may improve accuracy for other methods of detection, but may also be comparatively expensive (Markowitz and Bertagnolli 2009).

Higher delay costs, for both $b_1$ and $b_2$, imply higher optimal capacity. Alternatively, a higher marginal cost of screening capacity, $c$, implies lower optimal capacity.
3.5 Model Analysis: Choosing Average Time Between Screens for Fixed Capacity

In order to fully understand the relationship between the capacity planning decision and the average time until patients schedule the next screen, we now consider the social planner’s optimal decision for the time at home assuming the service capacity is fixed. We rewrite the social planner’s problem as follows.

\[
\min_{\tau : \mu > a_2(\tau)} TC = \lambda \left( \frac{\beta}{\alpha + \beta} \right) \left( \frac{\tau}{\tau + \alpha + \theta} \right) \left( \frac{b_1}{\tau + \alpha + \theta} + \frac{b_1}{\mu - \frac{\lambda(\tau + \beta + \alpha + \theta)}{(\alpha + \beta)(\tau + \alpha + \theta)}} \right) \\
+ \lambda \left( \frac{\beta}{\alpha + \beta} \right) \left( \frac{\theta}{\tau + \alpha + \theta} \right) \left( \frac{b_2}{\tau + \alpha + \theta} \right) + c\mu
\] (3.15)

In general it is difficult to show that the above function is unimodal in the decision variable, \( \tau \), where \( \frac{1}{\tau} \) is the average time at home. It is also difficult to explicitly solve the first order conditions in closed-form. We examine a special case of the model in order to gain some analytical insights and leave further analysis of the model for the Numerical Study.

3.5.1 Special Case, \( \alpha = 0; \theta = 0 \)

In this subsection, we assume that patients do not die from other causes, nor is cancer detected via other methods. Patients continue to be screened until cancer is eventually detected via screening. This model is an approximation but provides us with mathematical expressions which are helpful for obtaining insights. For this special case we evaluate the objective function at \( \alpha = 0 \) and \( \theta = 0 \) and define \( h = \frac{1}{\tau} \).
\[
\min_{h: \mu > a_2(h)} TC = c\mu + \lambda b_1 \left( h + \frac{1}{\mu - \lambda - \frac{1}{\beta b_1}} \right)
\]  
(3.16)

In the following proposition we show that the cost function is convex in the average time spent at home, \( h \). For the special case, we assume \( \mu > \lambda \), for long-run stability of the queueing network.

**Proposition 3.2.** The long-run average total cost given in Equation (3.16) is convex in the average time spent at home, \( h \).

Using the above proposition we solve for the optimal time at home in the following Theorem.

**Theorem 3.2.** The optimal average time at home which minimizes the long-run average total cost is given by the following two scenarios:

1. If \( \lambda \geq \beta \), no feasible solution

2. If \( \beta > \lambda \), \( h^* = \frac{\lambda + \sqrt{\lambda \beta}}{\beta(\mu - \lambda)} \), \( W^* = \frac{\lambda + \sqrt{\lambda \beta}}{(\mu - \lambda)\sqrt{\lambda \beta}} \), \( TC = c\mu + \frac{\lambda b_1(\lambda + \sqrt{\lambda \beta})(\beta + \sqrt{\lambda \beta})}{\beta(\mu - \lambda)\sqrt{\lambda \beta}} \)

Though the above results apply to an extreme case, some useful insights can still be obtained for the non-trivial case where \( \beta > \lambda \). This condition is required for feasibility because patients only leave the system by developing cancer; therefore current patients must develop cancer at least as fast as new patients enter the system or else the system is unstable (i.e. \( \frac{1}{\beta} < \frac{1}{\lambda} \)) for this special case. This condition is unreasonable in practice since the mean time until cancer detection should be much lower on average than the mean inter-arrival time. However, the special case results still provide insights into the optimal average time at home. The comparative statics for this special case are summarized in Table 3.2.

In this model, the service capacity is assumed to be exogenous. We look at the sensitivity
Table 3.2: Sensitivity Analysis: Choose time at home for fixed capacity, $\alpha = 0; \theta = 0$

<table>
<thead>
<tr>
<th></th>
<th>$h^*$</th>
<th>$a^*_2$</th>
<th>$W^*$</th>
<th>$TC^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta$</td>
<td>decreases</td>
<td>decreases</td>
<td>decreases</td>
<td>decreases</td>
</tr>
<tr>
<td>$b_1$</td>
<td>no effect</td>
<td>no effect</td>
<td>no effect</td>
<td>increases</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>increases</td>
<td>increases</td>
<td>increases</td>
<td>increases</td>
</tr>
<tr>
<td>$\mu$</td>
<td>decreases</td>
<td>increases</td>
<td>decreases</td>
<td>decreases*</td>
</tr>
</tbody>
</table>

*assuming $c = 0$

analysis table to observe the impact of a change in capacity given all other parameters are held constant. When capacity increases, the clinic has more slack and can increase the screening frequency by lowering the average time spent at home. The increase in capacity increases the demand for screens, but the faster service speed effect dominates and average wait time decreases.

If the capacity is exogenous, increasing the new patient arrival rate will put increased load on the system. The clinic will increase the time patients spend at home in order to control the demand for screens. The input into screening still increases, as does the wait time and the total cost. This result has important implications given the current environment of growing demand for preventive services. If capacity for colonoscopy screening is relatively fixed due to limited resources (physicians, clinic space, budget for capacity expansion), wait times will continue to increase and practical expectations may need to be set regarding the clinics ability to effectively deliver screening according to the guidelines.

3.6 Model Analysis: Joint Optimization for Capacity and Time Between Screens

In this Section, we allow the provider to simultaneously choose its capacity and the screening frequency. We rewrite the provider’s problem as follows.
In general, it is difficult to show that the above function is unimodal in the two decision variables. We examine a special case of the model in order to gain some analytical insights and leave further analysis of the model for the numerical study in a later section.

### 3.6.1 Special Case, \( \alpha = 0; \theta = 0 \)

In this subsection, we assume that patients do not die from other causes, nor is cancer detected via other methods. Patients continue to be screened until cancer is eventually detected via screening. This model is an approximation but provides us with mathematical expressions which are helpful for obtaining insights. For this special case we evaluate the objective function at \( \alpha = 0 \) and \( \theta = 0 \) and let \( h = \frac{1}{\tau} \).

\[
\min_{\mu, h: \mu > a_2(h)} TC = c\mu + \lambda b_1 \left( h + \frac{1}{\mu - \lambda - \frac{\lambda}{m}} \right) \quad (3.18)
\]

In the following proposition we show that the cost function is jointly convex in the average time spent at home, \( h \) and the capacity decision, \( \mu \). We assume \( \mu > \lambda \), for long-run stability of the queueing network.

**Proposition 3.3.** The cost function in Equation (3.18) is jointly convex in the capacity, \( \mu \), and average time at home, \( h \).
Table 3.3: Sensitivity Analysis: Choose time at home and capacity, $\alpha = 0; \theta = 0$

<table>
<thead>
<tr>
<th>$h^*$</th>
<th>$\mu^*$</th>
<th>$W^*$</th>
<th>$N^*$</th>
<th>$TC^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta$ decreases</td>
<td>decreases</td>
<td>no effect</td>
<td>decreases</td>
<td>decreases</td>
</tr>
<tr>
<td>$b_1$ decreases</td>
<td>increases</td>
<td>decreases</td>
<td>increases</td>
<td>increases</td>
</tr>
<tr>
<td>$c$ increases</td>
<td>decreases</td>
<td>increases</td>
<td>decreases</td>
<td>increases</td>
</tr>
<tr>
<td>$\lambda$ no effect</td>
<td>increases</td>
<td>decreases</td>
<td>no effect</td>
<td>increases</td>
</tr>
</tbody>
</table>

In the following Theorem we derive the optimal capacity and average time at home decisions for the above optimization problem.

**Theorem 3.3.** The long-run average total cost minimizing capacity and average time at home decisions are given by $\mu^* = \lambda + \lambda \sqrt{\frac{b_1}{3c}} + \lambda \sqrt{\frac{b_1 \beta}{c}}$; $h^* = \sqrt{\frac{c}{b_1}}$; $W^* = \sqrt{\frac{c}{b_1 \lambda}}$; $TC = c\lambda + 2\lambda \sqrt{\frac{ck}{b_1}} + 2\sqrt{c b_1 \lambda}$

Though the above results apply to an extreme case, some useful insights can still be obtained.

The comparative statics for this special case are summarized in Table 3.3.

Using sensitivity analysis for this special case, we compare in greater detail the impact of a change in a parameter on the average time waiting at home, $h$, compared to the average time waiting for a screen once requested, $W = \frac{1}{\mu - a_2}$. When the new patient arrival rate, $\lambda$, increases, screening frequency does not change but the average appointment delay (wait time) decreases meaning greater appointment access for patients. The change in the new patient arrival rate has a greater (absolute) impact on the appointment delay than on the screening frequency. When the cancer rate, $\beta$, increases, patients request screens more frequently, but the expected appointment delay does not change. The change in the cancer rate has a greater (decreasing) impact on the screening frequency than on the appointment wait time. When the delay cost, $b_1$, increases, patients will screen more frequently and also spend less time waiting for a screening appointment. The change in time spent waiting to schedule (at home) is given by
\( \frac{\partial h^*_1}{\partial b_1} = -\frac{1}{2b_1} \sqrt{\frac{c}{\beta b_1}} \). The change in average wait time is given by \( \frac{\partial W^*_1}{\partial b_1} = -\frac{1}{2b_1} \sqrt{\frac{c}{\beta \lambda}} \). The change in the delay cost has a greater (absolute) impact on the time waiting to schedule if \( \beta < \lambda \) and a greater (absolute) impact on average appointment wait time if \( \beta > \lambda \).

A surprising result in this model is that as the cancer incidence rate increases, the capacity decreases. This is mostly due to the special case assumptions. Since patients only leave the system once cancer is detected through screening, and not from dying due to other causes, then a higher incidence rate means that patients leave the system faster and fewer screens are needed. The total cost of the system is also lower. Though the average time at home decreases and patients request screens more frequently while alive, the effect of an increased rate of system exit dominates, and fewer screens are requested per new arrival on average.

The sensitivity of the optimal decisions, objective function, and performance measures with respect to the cost parameters remains intuitive. Higher delay cost implies higher capacity, less time waiting for a screen once requested, less time spent at home between screens, and higher total cost. Higher capacity cost implies lower capacity, increased waiting time for a screen once requested, more time spent at home on average, and higher total cost.

### 3.7 Numerical Study

We conduct a numerical study in order to solve the model using parameters estimated from real-world data. The numerical study provides further insights into the more complicated models which cannot be solved analytically.
3.7.1 Estimating Parameters

First we describe how we estimate each of the parameters of the model using public health data for colorectal cancer.

- We assume a time unit of weeks.
- The current guidelines endorsed by the CDC and the USPSTF recommend routine screenings beginning at age 50 with a FOBT test every year, sigmoidoscopy every 5 years, and a colonoscopy every 10 years (CDC 2013b). We assume that patients enter the system at age 50 and wait a random time at home before their first screen. We assume that the time at home between age 50 and the first screen follows the same distribution as the time spent at home between any subsequent screens. Our base estimate for the average time (at home) until the next screen is scheduled is 520 weeks (10 years). While colonoscopy screening is recommended to begin at age 50, not all patients will immediately request a screen once turning age 50. We could adjust this parameter, by assuming that patients enter the model at age 40 or 45 so that a higher number have requested a screen by age 50.

- The death rate due to other causes, $\alpha = \frac{1}{1623.5284}$. From data collected from the Social Security Administration (SSA) 2009 period Actuarial Life Table we observe the remaining life expectancy at age 50 to be 29.35 years for males and 33.02 years for females (SSA - Office of the Chief Actuary 2013). Using US Census data, we calculate the population distribution between sexes to at age 50 to be approximately 51% female and 49% male (U. S. Census Bureau 2013). We use a weighted average on the two expected remaining lifetimes to calculate an average remaining life expectancy of 31.2217 years. The patients
average remaining lifetime when entering the system is given by an exponential distribution with mean 31.2217 years. We solve for alpha by taking the inverse of this mean converted to weeks. Since the death rate due to colorectal cancer is relatively small, we do not adjust the life expectancy to be only deaths due to other causes; therefore we slightly over-estimate the death rate due to other causes.

- The rate of people who reach screening age, $\lambda = 10.662$ patients per week. We observe data from the US Census Bureau on population distribution by age and we estimate a current rate of approximately 12,500 people in the US turn 50 every day (U. S. Census Bureau 2013). This amounts to 87,500 people per week. We divide this by an estimate of the number of clinics administering colonoscopies, 8207, to get the average number of patients per clinic who turn 50 per week (Seeff et al. 2004).

- The cancer incidence rate, $\beta = 0.00001924$. From data collected from the Center for Disease Control and Prevention we observe that 52.7 men per 100,000 men and 39.7 per 100,000 women are diagnosed with colorectal cancer each year (CDC 2013a). We average the incidence rates to get the incidence rate for all people assuming the population is 51% female and 49% male to get an incidence rate of 46.070 per 100,000 people (U. S. Census Bureau 2013). Using the data on number of people who turn 50 each week we solve the following equation for $\beta$ using the estimates below and the mortality rate calculated above. Note that using this equation to estimate $\beta$ assumes a stable population size. An alternative estimate from the National Cancer Institute of 102,480 new cases of colorectal cancer in 2013 or the estimate from the first paragraph of the introduction could also be used on the right hand side of the equation and will provide a slightly higher estimate (currently the right hand side is estimated to be 138,210 new cases per year). Our estimates
for $\alpha$ and $\beta$, imply that about 3% of all people will develop colorectal cancer at some point in their lifetime (though it may go undetected in some who die from other causes before detection).

\[ P(Cancer \text{ After } 50) = \frac{(Incidence \text{ Per } 100000)(USPopulation)}{100000} \]

\[ (Turn \ 50 \ Per \ Year) = (12500)(365) = 4,562,500 \]

\[ P(Cancer \text{ After } 50) = \frac{\beta}{\alpha+\beta} \]

\[ (Incidence \text{ Per } 100000) = 47.070 \]

\[ (USPopulation) = 300,000,000 \]

Substituting and solving for $\beta$ we get the above parameter estimate.

\[ (4,562,500) \left( \frac{\beta}{1623.5284 + \beta} \right) = \frac{(47.070)(3,000,000)}{100000} \]

The rate until cancer is detected via other methods $\theta = \frac{1}{\frac{1}{2}60}$. We found this parameter difficult to estimate, so we assume in the base case that the average time until cancer is detected via other methods once the person has cancer to be half of the established screening guideline.

The cost of colonoscopy capacity, $c = 3.081$. We observe an estimate for the average cost of a colonoscopy to be $3,081 which is the average cost of a colonoscopy estimated by BCBSNC for an uninsured patient ((CostHelper.com 2013)). This cost does not include the cost associated with lost utility or lost productivity when the patient actually attends a screen. We write the costs in terms of thousands of dollars per week. Rosenthal (2013) states the average price in the US as $1,185, but this appears to be an out of pocket price, which does not reflect the total cost and does not include additional fees for sedation.
• The cost of delay when cancer is detected via screening $b_1 = 0.3851$. This parameter is also difficult to estimate due to the slow growing nature of colorectal cancer and lack of available data. Luo et al. (2010) provide estimates of 1-year costs attributable to colon cancer by stage of diagnosis. They find that patients diagnosed with cancer in local stage had the lowest costs ($27,551), followed by patients with distant stage ($29,933), and patients with regional cancer had the highest cost ($30,748). From these findings we see that there is a high fixed cost for treating cancer from any stage (even local) and the increase in costs by stage above does not account for other costs such as increased chance of death from cancer or disutility for the patient. We assume that 8 weeks of delay for a patient with cancer will cost as much in treatment as one colonoscopy.

• The cost of delay when cancer is detected other methods $b_2 = 0.7703$. This parameter is also difficult to estimate so we assume it to be twice the delay cost when cancer is detected via screening.

• The base capacity is $\mu = 33.327$. We observe from Seeff et al. (2004) an estimate of 14.2 million colonoscopies for 8,207 practices in 2002. This translates to an average screening rate of 33.327 colonoscopies per week per clinic. Assuming that utilization for colonoscopies is very high (as observed by high waiting time) we use this estimate as the estimate for capacity. However, the authors also find that all physicians combined report that they could increase to 22.4 million colonoscopies within 1 year. This translates to a capacity of $\mu = 52.488$. We can use the first estimate as a low estimate for average weekly clinic capacity and the second estimate as a high estimate for average weekly clinic capacity.
3.7.2 Capacity Decision, Fixed Time at Home

In this subsection of the numerical study, we provide numerical analysis for the model described in Sections 3 and 4, where the average time at home is fixed and the provider chooses its capacity. This is an important model to consider because the guidelines for colorectal cancer are relatively well-established and considerable education efforts have already been executed. It may be difficult for policy makers and providers to greatly influence the average time patients wait before trying to schedule a screen. The proceeding analysis also addresses the question of how much capacity is needed to meet a specified adherence level or “current” adherence level as modeled by a random time at home before scheduling the next screen.

Table 3.4 provides numerical calculations of the capacity decision, equilibrium wait time, and optimal total cost for the base parameters across six scenarios which vary the delay costs within the set \( \{ [b_1 = \frac{c_1}{16}, b_2 = \frac{c_2}{8}], [b_1 = \frac{c_1}{8}, b_2 = \frac{c_2}{4}], [b_1 = \frac{c_1}{4}, b_2 = \frac{c_2}{2}] \} \) and the arrival rate within the set \( \{ \lambda = 10.662, \lambda = 42.647 \} \). The two arrival rates correspond to an average size clinic, and a clinic 4 times the average size. While it is difficult for us to estimate some of the parameters of the model such as the delay costs and the time until cancer is detected via other methods, our results indicate that the range for the optimal capacity at our base parameters falls below our estimate for the current average clinic capacity, i.e. \( \mu \in [32.454, 32.532] \leq \mu = 33.327 \). This would indicate that current average clinic capacity could be enough to meet all demand for cancer screening at the currently established guidelines. In the model, since colorectal cancer is fairly rare, the clinic keeps a very small capacity buffer (utilization close to 1), and the wait time for screening ranges from 6.413 to 12.825 weeks. In practice waits of 6-12 weeks could discourage patients from scheduling screens and be long enough to contribute to poor adherence.
Table 3.4: Optimal Capacity, Wait Time, and Total Cost

| λ = 10.662 | b₁ = \(\frac{c}{10}\), b₂ = \(\frac{c}{8}\) | 32.454 | 12.825 | 114.899 |
| λ = 10.662 | b₁ = \(\frac{c}{5}\), b₂ = \(\frac{c}{4}\) | 32.487 | 9.069 | 129.766 |
| λ = 10.662 | b₁ = \(\frac{c}{4}\), b₂ = \(\frac{c}{2}\) | 32.532 | 6.413 | 159.382 |
| λ = 42.647 | b₁ = \(\frac{c}{10}\), b₂ = \(\frac{c}{8}\) | 129.661 | 6.413 | 458.636 |
| λ = 42.647 | b₁ = \(\frac{c}{5}\), b₂ = \(\frac{c}{4}\) | 129.726 | 4.534 | 517.704 |
| λ = 42.647 | b₁ = \(\frac{c}{4}\), b₂ = \(\frac{c}{2}\) | 129.817 | 3.206 | 635.606 |

\[\alpha = 0.000616, \beta = 0.00001924, c = 3.081, b₁ = 0.385125, b₂ = 0.77025, \tau = 1/520, \theta = 1/260\]

to screening guidelines.

Using our model and base parameter estimates for average clinic capacity and new patient arrival rate, we estimate an average wait time of 1.11 weeks for screening at a clinic with current average capacity. The probability of waiting more than 2 weeks is 16.618%. In order to guarantee an average wait-time of less than \(t\) weeks the clinic would need \(W = \frac{1}{\mu-a₂} \leq t\). Given our assumption that cancer does not develop while waiting, the input rate \(a₂ = \frac{\lambda\tau(\tau+\alpha+\beta)}{\alpha+\beta}\) does not depend upon the capacity decision, \(\mu\). For our base set of parameters we estimate \(a₂ = 32.37\). Solving for \(\mu\) we obtain \(\mu \geq \frac{1}{t} + a₂\). Given our model assumptions we know the distribution of the wait time function to be exponential. In order to guarantee that no more than \(x\)% of patients wait more than \(t\) weeks for an appointment, we solve \(P(W > t) = x \geq e^{-(\mu-a₂)t}\).

To guarantee a service level of \((1-x)\)% of patients wait less than \(t\) weeks, the clinic requires a capacity of \(\mu \geq a₂ - \frac{\ln(x)}{t}\). In Table 3.5 we solve these two equations for different combinations of the threshold for time waiting, \(t\), and the percent waiting more than threshold, \(x\). We also show the percent change in total cost over the optimal for delivering the corresponding service level \(\%\triangle TC = \frac{TC-TC^*}{TC^*}\).

We graph the sensitivity of the four results from Table 3.1 which were not monotone to see whether these statics are monotone near our base case estimates. In Figure 3.3, we can see
Table 3.5: Capacity Needed to Guarantee Low Wait Times

<table>
<thead>
<tr>
<th></th>
<th>t=2</th>
<th>t=1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Avg.</td>
<td>x = 0.1</td>
</tr>
<tr>
<td>μ</td>
<td>32.876</td>
<td>33.528</td>
</tr>
<tr>
<td>%ΔTC</td>
<td>0.72%</td>
<td>2.23%</td>
</tr>
</tbody>
</table>

[α = 0.000616, β = 0.00001924, c = 3.081, b₁ = 0.385125, b₂ = 0.77025, τ = 1/520, θ = 1/260, λ = 10.662]

that while μ* is not generally monotone in the cancer rate, the optimal capacity decreases in β within the reasonable range of 0 ≤ β ≤ α.

In Figure 3.4, we see that the optimal cost is increasing in the cancer rate β for our base set of parameters though the cost function is not generally monotone in β. The patients who develop cancer will develop cancer faster and, on average, experience a longer delay before being screened. Patients also leave the system faster as they are now more likely to develop cancer sooner. When β increases, the patients who get cancer are now more likely to get cancer sooner in their life and the number of screens per new patient arrival decreases. The provider will decrease its capacity because demand is lower and the cost of capacity dominates the cost of delay.

In Figure 3.5, we see that the cost is not monotone in the rate until cancer is detected via other methods. In this case, if θ is extremely low, lower than α even, (meaning that on average the patient’s remaining lifetime will be shorter than the time it takes to detect the disease from other methods), then increasing θ will increase the total costs. Since θ is very small, increasing it slightly will have a significant impact on the average time until cancer is detected via other methods. An extremely rare event becomes less rare, and the higher delay cost will be incurred for those patients. If θ increases above a threshold, further increases lead to a decrease in total cost. This result gives insight into other forms of screening and diagnosis used in practice. In our
Figure 3.3: Optimal Capacity Increases in Mean Time Until Onset of Cancer

\[ \alpha = 0.000616, c = 3.081, b_1 = 0.385125, b_2 = 0.77025, \tau = 1/520, \theta = 1/260, \lambda = 10.662, b_1 = \frac{c}{\theta}, b_2 = \frac{c}{\tau} \]
model we assume these other forms of diagnosis do not require patients to schedule a follow-up screen, so faster diagnosis via other methods reduces the number of screens demanded and the lower capacity costs dominates the increased cost due to detection from other causes as opposed to detection from screening.

In Figure 3.6, we see that the cost function is increasing monotonically in the rate at home (decreasing in the average time at home) when the other parameters are at the base estimates. As shown analytically in Section 4, this is not generally the case, as the optimal cost function can be non-monotonic in the average time at home. Less time at home on average means increased screening frequency, and this means an increase in the number of screens demanded. For the base set of parameters, the higher capacity cost dominates the cost saved from shorter detection delay and optimal costs increase when the screening frequency increases.
\[ \alpha = 0.000616, \beta = 0.00001924, c = 3.081, b_1 = 0.385125, b_2 = 0.77025, \tau = 1/520, \lambda = 10.662 \]

### 3.7.3 Average Time at Home Decision, Fixed Capacity

In this subsection of the Numerical Study, we provide numerical analysis for the model described in Section 5, where capacity is limited and the provider can influence the average time patients spend at home. In this section we set \( \mu = 33.27 \). In Figure 3.7 we see that the long-run average cost function is convex in the inverse of the mean time at home within the feasible range. The upper boundary of the x-axis is given by the condition \( a_2(\tau) < \mu \). This condition guarantees that the total input for screening does not exceed the screening capacity and the queueing network is stable.

Table 3.6 provides numerical calculations of the average time at home decision, equilibrium wait time, and optimal total cost for the base parameters across six scenarios which vary the
Figure 3.6: Optimal Cost Decreases in Mean Time Until Next Screening Request

\[ \alpha = 0.000616, \beta = 0.00001924, c = 3.081, b_1 = 0.385125, b_2 = 0.77025, \tau = 1/520, \lambda = 10.662 \]

Delay costs within the set \{ \left[ b_1 = \frac{c}{5}, b_2 = \frac{c}{8} \right], \left[ b_1 = \frac{c}{5}, b_2 = \frac{c}{4} \right], \left[ b_1 = \frac{c}{5}, b_2 = \frac{c}{2} \right] \} and the arrival rate within the set \{ \lambda = 10.662, \lambda = 15.992 \}. We calculate the optimal decision by solving the first-order conditions numerically, checking the solution set for feasibility (non-negativity and queueing stability conditions) and then evaluating the total cost. The two arrival rates correspond to an average clinic market size, and a clinic 50% above average e.g. growth due to population aging, and/or increased insurance coverage. At the estimates for current average clinic capacity and clinic market size (number of people reaching age 50), we calculate the optimal average time at home to be within the range of \( h^* = [509.336, 513.883] \) based on our choices for the delay cost parameters. This estimate for optimal average time at home is close to the established guidelines of colonoscopy screening every ten years, but our current results under-estimate the demand for all colonoscopies by not including diagnostic colonoscopies or
Figure 3.7: Cost Function Is Convex in Mean Time at Home in Feasible Region

[\alpha = 0.000616, \beta = 0.00001924, c = 3.081, b_1 = 0.385125, b_2 = 0.77025, \mu = 33.273, \lambda = 10.662]
The clinic would prefer patients spend less time at home so that when cancer occurs it is more likely to be found via screening than via other methods. The wait time increases because the demand for screens has now increased for fixed capacity. The total time between screens, the sum of expected time at home and expected wait time, remains fairly constant (dropping only slightly). This effect is more pronounced when $\lambda$ is higher. Even under our assumption of no onset of cancer while waiting, higher delay cost due to other causes does not mean that the clinic merely shifts the lower time at home to more time waiting. Instead, there is some (small) change in the overall expected time between screens.

The second important sensitivity result from Table 3.6 is that increasing the new patient arrival rate by 50% for an average clinic implies that the optimal average time at home ranges from 763.801 weeks to 770.624 weeks. This result is significant, not because of its ordinal direction, but because of its magnitude and implication for current changes in the health care environment which many colonoscopy providers are facing. Increasing the demand by 50% implies that the optimal time at home increases approximately 255 weeks for our base set of parameters. Since the average clinic currently operates at high utilization, increasing the patient arrival rate by 50% implies an approximately 50% increase in the average time spent at home. The wait time increases very slightly (e.g. 3.071 to 3.329), especially compared to the large change in the optimal time at home between screens.

3.7.4 Joint Capacity and Average Time at Home Decisions

In this subsection of the Numerical Study, we provide numerical analysis for the model described in Section 6, where the provider may choose both the capacity and the average time patients spend at home. We calculate the optimal decision by solving the first-order conditions numeri-
Table 3.7: Optimal Capacity, Time at Home, Wait Time, and Total Cost - 1

<table>
<thead>
<tr>
<th>Scenario</th>
<th>$b_1 = c, b_2 = 2c$</th>
<th>$b_1 = 2c, b_2 = 4c$</th>
<th>$b_1 = 4c, b_2 = 8c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda = 10.662$</td>
<td>$\mu^* = 30.082$</td>
<td>$\mu^* = 64.402$</td>
<td>$\mu^* = 111.252$</td>
</tr>
<tr>
<td></td>
<td>$h^* = \frac{1}{\tau}$</td>
<td>$h^* = \frac{1}{\tau}$</td>
<td>$h^* = \frac{1}{\tau}$</td>
</tr>
<tr>
<td></td>
<td>$W^* = 565.395$</td>
<td>$W^* = 263.466$</td>
<td>$W^* = 152.338$</td>
</tr>
<tr>
<td></td>
<td>$N^* = 3.303$</td>
<td>$N^* = 1.835$</td>
<td>$N^* = 1.140$</td>
</tr>
<tr>
<td></td>
<td>$TC^* = 2.793$</td>
<td>$TC^* = 5.989$</td>
<td>$TC^* = 10.352$</td>
</tr>
<tr>
<td>$\lambda = 15.992$</td>
<td>$\mu^* = 45.123$</td>
<td>$\mu^* = 96.526$</td>
<td>$\mu^* = 166.706$</td>
</tr>
<tr>
<td></td>
<td>$h^* = \frac{1}{\tau}$</td>
<td>$h^* = \frac{1}{\tau}$</td>
<td>$h^* = \frac{1}{\tau}$</td>
</tr>
<tr>
<td></td>
<td>$W^* = 564.332$</td>
<td>$W^* = 263.266$</td>
<td>$W^* = 152.274$</td>
</tr>
<tr>
<td></td>
<td>$N^* = 2.694$</td>
<td>$N^* = 1.498$</td>
<td>$N^* = 0.931$</td>
</tr>
<tr>
<td></td>
<td>$TC^* = 2.798$</td>
<td>$TC^* = 5.994$</td>
<td>$TC^* = 10.357$</td>
</tr>
</tbody>
</table>

[$\alpha = 0.000616, \beta = 0.00001924, c = 3.081, \theta = 1/260$]

cally, checking the solution set for feasibility (non-negativity and queueing stability conditions) and then evaluating the total cost for any candidate solutions.

Table 3.7 provides numerical calculations of the average time at home decision, equilibrium wait time, and optimal total cost for the base parameters across six scenarios which vary the delay costs within the set $\{[b_1 = c, b_2 = 2c], [b_1 = 2c, b_2 = 4c], [b_1 = 4c, b_2 = 8c]\}$ and the arrival rate within the set $\{\lambda = 10.662, \lambda = 15.992\}$. We initialize the parameter estimates for the delay costs to be higher than in the earlier models to ensure non-trivial, feasible solutions exist (i.e. the delay penalty is large enough s.t. $\tau > 0$ and $\mu > a_2(\tau)$). In Table 3.7 we vary the magnitude of the delay costs while keeping the ratio between the delay costs constant. The two arrival rates correspond to an average clinic market size and a clinic 50% above average. In the parameter setting of $\lambda = 10.662; b_1 = c; b_2 = 2c$, we see that the optimal capacity is slightly below the current clinic average we observed from our data ($\mu = 30.082$ compared to $\mu = 33.27$).

The average time at home is slightly above the established guidelines ($h^* = 565.39$ compared to $h^* = 520$). The estimated average wait time for this case is 3.303 weeks with an average total cost of $336,111$.

Patterns in the results from Table 3.7 provide insight into the sensitivity of the optimal solution to the parameters. First, we observe sensitivity with respect to the arrival rate. When
the delay costs are $b_1 = c, b_2 = 2c$, a 50% increase in the arrival rate indicates a 50% increase in the capacity, the average time at home drops 0.19% and the average wait time drops by 18.44%.

When delay costs are $b_1 = 2c, b_2 = 4c$, the increase in capacity is 49.88%, the average time at home drops 0.08%, and the average wait time drops by 18.37%. When the delay costs are $b_1 = 4c, b_2 = 8c$ the increase in capacity is 49.85%, the average time at home drops 0.04%, and the average wait time drops by 18.33%. Across multiple settings, a 50% increase in the arrival rate leads to roughly a 50% increase in capacity, but as delay costs are higher, the percent increase in capacity due to a change in the arrival rate decreases slightly. The average time at home decreases only slightly due to a 50% increase in the arrival rate and this increase is diminishing as the delay cost increases. The wait time decreases by approximately 18% due to a 50% increase in the arrival rate and this decrease is slightly smaller when the delay costs are higher. These numerical results are fairly consistent with the intuition derived from the analytical results for the special case of $\alpha = 0; \theta = 0$. As the arrival rate increases, the capacity increases and the expected wait time drops, with little impact on the average time at home.

Second, we observe sensitivity with respect to the delay costs. When the new patient arrival rate is equal to 10.662, increasing the delay costs from $b_1 = c, b_2 = 2c$ to $b_1 = 2c, b_2 = 4c$ increases the capacity by 114.09%, decreases the average time at home by 53.40%, and decreases the average wait time by 44.44%. Increasing the delay costs from $b_1 = 2c, b_2 = 4c$ to $b_1 = 4c, b_2 = 8c$ increases the capacity by 72.75%, decreases the average time at home by 42.18%, and decreases the average wait time time by 37.87%. When the new patient arrival rate is equal to 15.992, increasing the delay costs from $b_1 = c, b_2 = 2c$ to $b_1 = 2c, b_2 = 4c$ increases the capacity by 113.92%, decreases the average time at home by 53.35%, and decreases the average wait time by 44.39%. Increasing the delay costs from $b_1 = 2c, b_2 = 4c$ to $b_1 = 4c, b_2 = 8c$ increases the
capacity by 72.71%, decreases the average time at home by 42.16%, and decreases the average wait time by 37.85%. For both arrival rates, the absolute decrease in the time spent at home due to delay costs is much greater than that for the expected wait time. This agrees with our earlier analytical result for the special case of $\alpha = 0; \theta = 0$ that increasing the delay cost will have a greater negative impact on the average time at home than on the expected wait-time if $\beta < \lambda$. In this case the cancer rate, $\beta$, is much less than $\lambda$, and we see that increasing the delay cost decreases the average time at home much more than the average time waiting. The numerical results above show that this argument holds true even when converting to percent changes. The percent decrease in time at home is also larger (more negative) in all cases than the percent decrease in average waiting time. The percent changes in capacity, average time at home, and wait time appear to be fairly consistent across both arrival rates.

In Table 3.8, we calculate the optimal solution for multiple combinations of $\beta$ and $\lambda$ with the other parameters set to base estimates. We observe the sensitivity of the parameters with respect to changes in the cancer rate, $\beta$. For both new patient arrival rates, $\lambda = 10.662$ and $\lambda = 15.992$, we see that increasing the cancer rate increases the optimal capacity, decreases the optimal average time at home, and decreases the optimal wait time. When $\lambda = 10.662$, increasing the cancer rate by 100% from $\beta = 0.00001924$ to $\beta = 0.00003848$ causes the capacity to increase by 108.05%, the average time at home to decrease by 53.35%, and the average wait time to decrease by 43.60%. When $\lambda = 10.662$, increasing the cancer rate by 100% from $\beta = 0.00003848$ to $\beta = 0.00007697$ causes the capacity to increase by 63.59%, the average time at home to decrease by 42.16%, and the average wait time to decrease by 36.07%. When $\lambda = 15.992$, increasing the cancer rate by 100% from $\beta = 0.00001924$ to $\beta = 0.00003848$ causes the capacity to increase by 107.89%, the average time at home to decrease by 53.29%, and the
average wait time to decrease by 43.54%. When \( \lambda = 15.992 \), increasing the cancer rate by 100% from \( \beta = 0.00003848 \) to \( \beta = 0.00007697 \) causes the capacity to increase by 63.55%, the average time at home to decrease by 42.14%, and the average wait time to decrease by 36.03%. These results differ from the analytical results we derived for the special case \( \alpha = 0; \theta = 0 \) in that an increased cancer rate does lead to lower wait times (in the special case there was no effect on wait time). The results are consistent with the special case, however, in that the impact of increasing the cancer incidence rate is greater on the average time at home than on the average wait time, both in terms of absolute and percent change.

In Table 3.9 we allow the parameter for the time until detection from other methods, \( \theta \), to vary and also vary the ratio of the delay cost parameters \( \{b_2 = 2b_1, b_2 = 3b_1\} \). Across both combinations of delay cost parameters, we observe that decreasing the time until cancer is detected via other methods leads to lower capacity, higher average waiting time at home, and higher average waiting time for the primary screen. When \( b_1 = c, b_2 = 2c \), and \( \theta = \frac{1}{260} \), increasing the mean time until detection via other methods by 1.5 years (78 weeks), leads to an increase of 8.788 units of capacity (29.21% increase), a decrease of 128.162 average weeks at home (22.67% decrease) and a decrease of 0.486 weeks waiting (14.71%). When \( b_1 = c, b_2 = 2c \), and \( \theta = \frac{1}{338} \), increasing the mean time until detection via other methods by 1.5 years (78 weeks), leads to an increase of 4.922 units of capacity (12.66% increase), a decrease of 99.207 average
weeks at home (22.69% decrease) and a decrease of 0.224 weeks waiting (7.95%). When \( b_1 = c, \)
\( b_2 = 2c \) and \( \theta = \frac{1}{260}, \) increasing the mean time until detection via other methods by 1.5 years
(78 weeks), leads to an increase of 6.643 units of capacity (13.12% increase), a decrease of 38.855
average weeks at home (11.60% decrease) and a decrease of 0.255 weeks waiting (9.18%). When
\( b_1 = c, b_2 = 2c \), and \( \theta = \frac{1}{338}, \) increasing the mean time until detection via other methods by 1.5
years (78 weeks), leads to an increase of 3.356 units of capacity (5.86% increase), a decrease of
16.356 average weeks at home (5.53% decrease) and a decrease of 0.134 weeks waiting (5.31%).

When the delay cost ratio is higher, increasing the time until cancer is detected by other methods
will have a greater (absolute) impact on the optimal solution. Also, we see that increasing the
time until cancer is detected via other methods appears to have a greater impact on average
time at home (both absolute and percentage) than it does on average waiting time.

Additionally, we see from Table 3.9 that increasing the detection delay cost ratio so that
the cost of detection via other methods becomes much more costly than detection via screening
generally leads to higher optimal capacity, lower average time at home, and lower expected
wait times. When \( \theta = \frac{1}{416}, \) increasing the delay cost ratio from 2 to 3 leads to an increase of
16.821 units of capacity (38.41%), a decrease of 58.368 average weeks at home (17.27%) and
a decrease of 0.203 (7.83%) average weeks waiting. When \( \theta = \frac{1}{338}, \) increasing the delay cost
ratio from 2 to 3 leads to an increase of 18.387 units of capacity (47.30%), a decrease of 141.219
average weeks at home (32.30%) and a decrease of 0.293 average weeks waiting (10.40%). When
\( \theta = \frac{1}{260}, \) increasing the delay cost ratio from 2 to 3 leads to an increase of 20.532 units of
capacity (68.25%), a decrease of 230.526 average weeks at home (40.77%) and a decrease of
0.524 average weeks waiting (15.86%). Therefore, when the average time until detection from
other methods is lower, increasing the delay cost ratio appears to have a greater impact on the
optimal solution. We see that changing this ratio also appears to have a greater impact on average time at home (both absolute and percentage) than it does on average wait time.

### 3.8 Conclusions and Future Work

We develop a queueing network model to analyze capacity planning and screening guideline decisions for a social planner who minimizes the long-run average total of detection delay cost and capacity cost. Patients spend a random amount of time at home before attempting to schedule a screen for which they must wait according to the queueing dynamics at the endoscopy suite. Between any two screens, patients may either die from other causes or develop cancer. If cancer is developed, it may be detected via a screen or via other methods. We solve for the optimal capacity decision when the average time patients spend at home is exogenous, and we perform sensitivity analysis on the optimal solution. When the capacity is fixed and the provider optimizes the average time at home, the function is not convex in the general case so we solve a special case and perform sensitivity analysis. We investigate the general case via numerical study. When we allow the provider to jointly optimize both the capacity decision and the average time at home, we consider a special case for our analytical results because it is difficult to show unimodality and to solve the model in the general case. We use a numerical
study to generate real-world estimates of our results from public health data and to further investigate the model under general assumptions.

A primary objective and contribution for this research, beyond developing a model to optimize and guide capacity planning decisions for colorectal cancer screening, is capturing and understanding the tradeoff between screening guidelines and screening capacity. If screening guidelines are developed using medical decision making which only takes into account the utility and health state tradeoffs for a single individual, they may suggest policies which are unsustainable or sub-optimal given the system capacity or cost of adding additional capacity. If we assume the average time at home to be exogenous, e.g. due to well-established guidelines, we find that increasing the average time at home causes the optimal capacity to increase due to increased demand for screens, but also an increased capacity buffer. The increased buffer results in lower average waiting time. The impact on total cost is non-monotonic which leads us to the joint optimization model where we find through our numerical study that changes in the parameters which cause the optimal capacity to increase generally correspond to decreases in the optimal average time at home (and often the average wait time as well). An exception to this general pattern is the special case where patients only leave the system by getting cancer. In this case increasing the cancer rate would lead to lower capacity and lower average time at home as the provider wants to detect cancer faster, but patients also leave the system sooner. In reality, patients are much more likely to die from other causes than colorectal cancer, so this exception seems unlikely in practice.

From our analysis, we obtain several important insights for policy makers and cancer screening providers. Our findings suggest that if providers have the flexibility to adjust capacity at fairly constant marginal cost, increased demand will lead to improved service delivery in terms
of reduced wait time for colonoscopy. We find that as the new patient arrival rate increases, e.g. due to shifts in the population distribution or increased insurance coverage, the optimal capacity increases due to an increased demand for screens but also an increased capacity buffer. If the new patient arrival rate increases approximately 400% at an average clinic, we see that the capacity increases approximately 399%. This is because the clinic is operating at very high utilization and therefore keeps a relatively low slack (capacity - input rate) compared to capacity itself. The increase in the capacity buffer leads to lower average wait times. If clinics are capacity constrained, and choose only the average time at home, we see from our numerical study that increasing the new patient arrival rate by 50% leads to a roughly 50% increase in the average time spent at home with little effect on wait time. The social planner prefers that patients screen less frequently (wait longer on average until scheduling the next screen) when they are capacity constrained. This lengthens the effective screening guidelines to what the provider can reasonably deliver to the population. In reality, another way the shortage of capacity may play out is through increased wait-times for appointments rather than by adjustments in the time until the patient tries to schedule her next screen. Some patients may become frustrated with how lengthy appointment delays make it more challenging to adhere to guidelines and those patients may become less likely to be compliant. When social planners have the flexibility to adjust provider capacity and influence the average time patients spend at home, increasing the new patient arrival rate will again lead to lower wait times with little impact on the average time at home. This result is supported by the range of parameters estimated for our numerical study from public health data, and also derived analytically for a special case of the parameters. These results indicate that while growing demand is a concern in capacity-constrained environments, it can also result in improved appointment access when resources are available to adjust
capacity accordingly.

A challenge with cancer screening, and colorectal cancer in particular, is asymptomatic disease progression and the accuracy of affordable alternatives for screening and diagnosis such as FOBT. In our model we find that if the time until detection from other methods decreases, i.e. patients are encouraged to use FOBT tests with greater frequency and improved accuracy, this will reduce the optimal colonoscopy screening capacity required and likely decrease the optimal screening cost. Though the optimal cost is not generally monotone in the time at home, we observe it to only be increasing for a range of parameters which is unlikely in practice (e.g. average time until detection from other methods exceeds average lifetime). Lower screening capacity will lead to higher expected wait time for those who do request screens. In our numerical study of the joint optimization model, we also see that decreasing the time until detection via other methods will result in lower optimal capacity, higher optimal average waiting times, higher optimal average time at home, and lower optimal average cost. The provider prefers that patients wait at home longer so that the disease is more likely to be detected via other methods before requesting an expensive screen. An exception to these findings in practice is that colonoscopy is often used as a follow-up screen to confirm suspected diagnosis from other methods (e.g. positive FOBT test or symptoms). The assumptions of our model could be adjusted to account for such a setting. It could be in that case that faster detection from other methods implies an increase in capacity along with higher wait times and either higher or lower costs. Another caveat is how these results should be interpreted across different types of “other detection.” For example, if cancer is being detected faster because it is spreading to other organs and parts of the body faster, then a correlation between lower time until detection from other causes and higher detection delay penalty for other detection may exist. Future research could examine the
effect of such a correlation or model cancer detected via other screens (e.g. FOBT) separately from cancer detected via late stage progression, symptoms, or death.

Using public health data collected for our numerical study, our analysis also provides estimates of current capacity needs, economic validation of guidelines, and service delivery performance measures. Assuming the average time at home to be fixed at the current guideline of 10 years (520 weeks), we use public health data to estimate our model parameters and estimate that the average clinic sees a market size of 10.662 new patients per week and should provide an optimal capacity of 32.487 screens per week at an estimated long-run average weekly cost of $129,766 (capacity operating cost + delay cost). The average wait time will be 9.069 weeks. In order to guarantee that no more than 5% of patients are delayed more than 2 weeks for screening, the provider should choose a weekly capacity of 33.874 which results in an increase in the long-run average weekly costs of of 3.05% over optimal. While these estimates provide some insight into screening capacity needs, the model we develop is a stylized optimization model which was built for theoretical analysis as opposed to empirical estimation. Certain parameters of our model, such as the delay costs, are difficult to estimate due to the more implicit nature of what they represent. Initializing the model with data which is more specific to a particular market or geographic region would increase the reliability of the model estimates. We leave for future research further empirical estimation of colonoscopy screening capacity needs.

There are several opportunities for future research stemming from our model. One primary limitation of the model is that the exponential distribution assumptions may not fully characterize consumer behavior. The time patients spend at home might be more correctly modeled as a deterministic time (e.g. 10 years) plus some random time until the patient remembers to call and make the appointment or is reminded to do so by a primary care provider. We leave
this issue for future research, because while it may accurately describe the real-world behavior, the added complexity to the model may have little marginal benefit in terms of the research questions posed in this paper. This paper primarily focuses on issues regarding the design of the queueing network; whereas future research may emphasize problems regarding queueing control such as how to prioritize screens based on risk factors and how long it has been since the patient was last screened. Other future research might investigate reasons for poor adherence rates for colonoscopy screening and consider to what degree poor operational issues such as appointment access are drivers for poor patient adherence to screening guidelines. Other work might use location network analysis to investigate issues relevant to state and federal policy makers such as where to invest in additional cancer screening resources and where to build new endoscopy suites. Finally, future research might extend the general approach used within this work to other types of cancer or other diseases. Some of the assumptions of this paper, and most of the parameter estimates for the numerical study, were specifically motivated by the application to colorectal cancer; however the model is generalizable to screening for other diseases. Models which investigate the tradeoffs between supply planning and provider-driven demand in health services are beginning to be investigated more by the operations management and operations research community, and future work could bring about greater understanding of these types of problems.
Chapter 4

Revenue Management for Outpatient Appointments

4.1 Introduction

Outpatient clinics use appointment systems to match supply and demand while balancing goals of provider service efficiency with timely access for patients. Appointment scheduling helps to smooth workflow for resources and reduce variability in patient wait time for service. However, poor appointment utilization and excessive delays for outpatient appointments are widely recognized as significant barriers to effective health care delivery. Several studies document the high financial costs associated with no-shows and cancellations as well as their impact on staff satisfaction and productivity. Moore et al. (2001) estimate the cost of no-shows and cancellations at 3% to 14% of revenues. Other studies emphasize the cost of long appointment delay, the time between the request for and service of an appointment, highlighting lower patient satisfaction, decreased quality of care, and system-wide costs incurred when patients end up in the emergency room for non-emergency treatment. Strunk and Cunningham (2002) conclude that the percent of patients reporting an inability to obtain a timely appointment rose from 23% to 33% from 1997 to 2001.
Much evidence suggests a relationship between appointment waste and appointment delay. The longer the appointment delay, the less likely the patient is to attend the appointment, and in turn, wasted appointments, due to no-shows and late cancellations, do nothing to reduce the backlog of patients waiting. Festinger et al. (2002) randomly assign outpatient clients to different appointment delays and find that 72% of subjects scheduled 1 day later attend their appointments compared to 41% for 3 days later and 38% for 7 days later. Gallucci et al. (2005) find a similar relationship between appointment delay and rate of kept appointments for referrals to an outpatient program at a community mental health center. Figure 4.1 demonstrates the relationship between appointment waste and delay across three types of clinics within a major university health care system. Clinics, therefore, have incentive to reserve capacity for patients who schedule closer to the appointment date as they more likely to attend the appointment. In this study, we analyze how clinics should allocate appointments to customer classes with heterogeneous no-show rates by developing a capacity control and overbooking model for outpatient appointments.

Many proposed solutions for better outpatient appointment scheduling have seen limited
success in practice because they emphasize either appointment waste or appointment delay without considering the relationship between them. Solutions which emphasize service efficiency include no-show penalties, appointment reminders, and overbooking. In practice, it is difficult to collect financial penalties from patients who do not attend appointments. Other solutions emphasize improving customer access by dedicating capacity or resources to urgent requests. Clinics may dedicate one resource to urgent requests such as a “same-day doctor,” and larger health care systems may administer urgent care clinics or “minute clinics”. Healthcare providers may also “carve-out” (reserve) capacity for urgent requests. Problems with these solutions are high fixed costs associated with operating a resource dedicated to urgent problems and the resource costs associated with triaging which patients qualify to use capacity reserved for urgent requests. Clinics, such as private boutique clinics, may also reduce delay by limiting panel size, or new patients, as a means of controlling demand upfront.

Open access, also known as same-day scheduling or advanced access, is a popular paradigm in which the clinic attempts to “do today’s work today” by offering each patient a same-day appointment and/or encouraging patients to book within a short window (Murray and Tantau 2000, Murray and Berwick 2003). Open access reduces the need for triage, by emphasizing patient-centered care as a means for simultaneously achieving lower appointment waste and delay simultaneously. Open access alleviates some problems, but if not managed effectively, can lead to greater variation in daily demand and conflicts with patient preferences. A pure open access system, similar to a walk-in clinic where no advance appointments are allowed, could make the clinic more vulnerable to costly under- or over-utilization because of the absence of an appointment system buffer. These costs can be particularly high in clinics with providers who are not in clinic full-time, such as academic medical centers with residents and faculty. Open
access may also conflict with preferences of patients who must make advance arrangements such as transportation, absence from work, childcare, or lodging. The underlying issue of managing clinic capacity is not eliminated with open access. Clinics still must decide how to allocate appointments to different demand classes in a way that is both patient-centered and efficient. While open access is generally presented as an alternative to traditional advance scheduling, it is unclear exactly how the two may be effectively used together as a hybrid system.

How much fixed service capacity to allocate to different demand classes and whether to overbook are problems faced in outpatient appointment scheduling, but they are common to many service industries. Within operations and management sciences, these problems are typically referred to as capacity control decisions which fall under the broader field of revenue management. When there are advance purchase and reservation options, firms must also decide whether to overbook their fixed capacity to buffer against no-shows and late cancellations. Demand classes are customarily differentiated by features such as price and service package offered; however, differences in no-show and cancellation rates among demand classes are particularly significant in outpatient clinics which are limited in their ability to collect upfront deposits for service or to re-sell a wasted reservation to a late-arriving customer. Frequently, capacity allocation and overbooking decisions are considered separately, but several joint capacity-control and overbooking models, developed under the assumption that no-show probabilities are the same for all customers, show firms can achieve better results by integrating the two decisions (Talluri and van Ryzin 2005). We relax the assumption that no-show rates are homogeneous across demand classes, and show how firms can improve performance by accounting for these differences.

We construct a joint overbooking and capacity control model to study the optimal appointment allocation and overbooking decisions for outpatient clinics when patient demand classes
differ with respect to no-show rates. In Section 2, we review the literature on joint overbooking and capacity control and the literature on outpatient scheduling. In Section 3 we formulate a two-stage joint overbooking and capacity control model and provide structural results regarding the optimal policy. Under general assumptions, a closed form expression for the optimal policy can be difficult to derive; therefore, we develop lower and upper bounds and near-optimal approximations. In Section 4 we conduct a numerical study to compare the performance of the optimal policy with bounds and approximations, as well as policies from previous literature and clinic practice. In Section 5 we provide conclusions and insights from our work.

4.2 Literature Review

Our work is positioned at the intersection of the revenue management literature addressing joint capacity control and overbooking and the outpatient appointment scheduling literature. Talluri and van Ryzin (2005) provide a comprehensive review of the revenue management literature. They consider capacity control and overbooking separately prior to integration. The simple two-class allocation rule proposed by Littlewood (1972), is a seminal concept in traditional capacity control models. Assuming both a discount and a full fare segment, with independent random demands where the full fare demand, $D_1$, arrives later, Littlewood’s rule specifies that with $x$ units of capacity remaining, a discount unit should be sold as long as the discount fare, $p_2$, equals or exceeds the product of the full fare price, $p_1$, and the probability that full fare demand is $x$ or higher, i.e. $p_1 P(D_1 > x) \leq p_2$. Subsequent work applies stochastic dynamic programming to analyze models with more than two demand classes and dynamic models with simultaneous booking - e.g. Belobaba (1989).

Traditional overbooking models use both statistical service levels and economic criteria.
The number of customers who arrive for service is often modeled as a binomial random variable, \( Z(x) \), which depends on a static show-up probability, \( q \in [0, 1] \), and the number booked, \( x \). Economic criteria generally assume a cost of denying service to an overbooked customer, \( h \), when service arrivals exceed capacity. This may be a direct cost, such as a compensation voucher for future service, or an indirect cost such as loss of goodwill. The optimal solution for the static overbooking model is to book up to \( x \) so long as
\[
P(Z(x-1) > C-1) \leq \frac{p}{hq}
\]
where \( p \) denotes the revenue per customer (collected at time of booking) and \( C \) denotes capacity.

Dynamic overbooking models allow booking limit adjustment after realization of cancellations prior to date of service.

For joint capacity control and overbooking problems, Talluri and van Ryzin (2005) provide four assumptions for incorporating no-shows and cancellations into the single-resource capacity control model: (1) Cancellation and no-show probabilities are the same for all customers; (2) Cancellations and no-shows are independent across customers; (3) Cancellations and no-shows in any period are independent of the time the reservations on hand were accepted; (4) Refunds and denied service costs are the same for all customers. Various authors, e.g. Subramanian et al. (1999), study optimal booking limits under those four assumptions. We relax the first assumption with respect to no-show rates.

Previous literature on outpatient scheduling can be grouped into three significant streams: healthcare practice management research, outpatient appointment scheduling simulation, and analytical models for appointment scheduling. Healthcare practice management literature summarizes the challenges in outpatient appointment scheduling, approaches for implementing open access, as well as case studies of scheduling policies across a variety of clinic settings. The term open access is commonly credited to Murray and Tantau (2000). Murray and Berwick (2003)
further developed the concept under the name “advanced access” providing six elements for effective implementation. Numerous authors have since reported on both the improvements and challenges encountered when implementing open access within various clinical settings (e.g. Solberg et al. (2004), Belardi et al. (2004), Bundy et al. (2005), Dixon et al. (2006)). Other authors, such as Lamb (2002) and Mehrotra et al. (2008) have cautioned against open access, contending that it leaves clinics susceptible to daily demand variation. Gupta and Denton (2008) point out three challenges of implementing open access: (1) daily demand variation and spillover demand; (2) forecasting patient preferences from truncated data; (3) variation between clinic resources with respect to panel size and availability.

Discrete-event simulation is often used to analyze complex appointment scheduling systems with features such as multiple booking attempts or follow-up appointments. Cayirli and Veral (2003) and Denton and Gupta (2003) provide extensive reviews of the literature on outpatient appointments. Many studies analyze in-clinic waiting time and physician utilization as opposed to delay for appointments, although more recent work addresses open access or appointment delays specifically. Giachetti et al. (2005) find that open access is a viable strategy for reducing clinic throughput time. Kopach et al. (2007) find that correct configuration of open access can significantly improve clinic throughput rates. Giachetti (2008) find that eliminating multiple appointment types and segregating repeat no-show patients can effectively reduce appointment delay.

Analytical models of outpatient scheduling generally emphasize either indirect appointment wait (appointment delay) or direct appointment wait (in-clinic wait time). Cayirli and Veral (2003) and Gupta and Denton (2008) survey this body of literature and observe an emphasis on patient scheduling and resource utilization on the appointment date (in-clinic wait time).
In the direct-wait models, the decision variables typically include the number and length of appointment slots, the number of patients to schedule in each slot and the service priority (FCFS or by appointment time). Several such papers allow for overbooking by considering a static no-show probability, e.g. Kaandorp and Koole (2007), Kim and Giachetti (2006), LaGanga and Lawrence (2007), Robinson and Chen (2010), and Cayirli et al. (2006). Muthuraman and Lawley (2008) develop a sequential scheduling model with multiple no-show probabilities and exponential service which maximizes daily profits comprised of service revenues, patient waiting costs, and staff overtime costs. The sequential model implies that patient requests arrive individually, and the clinic must schedule the patient to a slot or terminate scheduling. The authors show that scheduling should be terminated according to an optimal stopping rule and develop a myopic scheduling algorithm to determine the best slot for each patient. Zeng et al. (2009) extend the work of Muthuraman and Lawley (2008) by emphasizing class-dependent no-show rates and deriving properties of the optimal sequential schedule which were not previously characterized. Chakraborty et al. (2010) also extend the work of Muthuraman and Lawley (2008) by considering general service distributions, showing that the unimodality of the profit function and resulting stopping rule are independent of the service distribution.

Other recent articles examine indirect appointment wait time. Green and Savin (2008) present a queueing model to optimize panel size and provide timely access to outpatient care. Assuming patients prefer the first available appointment, the authors obtain expressions for optimal panel sizes for M/D/1/K and M/M/1/K queues with backlog dependent cancellations and no-shows. These results provide upper and lower bounds, respectively, for simulation results which incorporate more realistic assumptions on patient preferences. Liu et al. (2010) study dynamic heuristic policies for outpatient scheduling under time-dependent no-shows and
cancellations. The authors formulate an infinite horizon Markov decision process where a clinic chooses to schedule an arriving appointment request to any open booking date.

Several authors study the direct application of revenue management to healthcare settings. Gupta and Denton (2008) present two challenges with adapting revenue management models from other industries (such as airlines and hospitality) to fit healthcare. The first challenge is that the patient choice function has multiple elements such as provider, date, time of day. The second challenge is that differentiated prices are not used to control access to clinic capacity. Gupta and Wang (2008) address the first challenge by modeling patient choice in a primary care clinic. They prove the optimal policy is threshold-type for a single physician clinic under weak conditions on the choice behavior. For the multi-physician model, the authors partially characterize the structure of the optimal policy and provide policies which perform well in numerical tests based on clinic data. The authors assume two classes of patients, regular patients and same-day patients. All regular patients, those requesting an appointment in advance, pay the same price if booked and incur the same penalty if turned away, while same-day patients pay higher prices. They note that optimizing accept/deny decisions across multiple booking attempts has proven to be intractable; although in reality, patients may reattempt booking at a later stage or appointment date. Significantly, they do not allow for no-shows, cancellations or overbooking.

Our work contributes to the above literature by studying how outpatient clinics can use capacity control and overbooking to allocate appointments to customer classes with heterogeneous no-show rates. We need not assume patient classes differ by revenue as in previous work. We establish some structural characterizations of the optimal solution and compare our allocation policies with other policies commonly used in the literature and in practice. Gupta and Denton
(2008) point out that none of the previous overbooking models from outpatient scheduling have considered the link between appointment delay and no-show rates. More recent papers address time-dependent no-show rates, and we enhance this research by considering a revenue management application as a different type of patient demand management. We consider a model where a clinic may limit appointments allocated to an early arriving patient demand class in order to reserve appointments for a later class. Our approach contrasts with Muthuraman and Lawley (2008), Zeng et al. (2009), and Chakraborty et al. (2010) who suggest ordering patients and require scheduling to be terminated for all patients simultaneously. Compared to Green and Savin (2008), appointment allocation provides the clinic with greater flexibility than controlling its panel size. Instead of assigning patients to a range of appointment dates as with Liu et al. (2010), our model assumes patients prefer a single date, thereby keeping with the spirit of “patient-centered scheduling.” Our model, like much of the prior research, shows that no-show heterogeneity makes the problem complex and often intractable, requiring the formulation of heuristics and approximations.

4.3 Model

4.3.1 Setting and Notation

We consider a two-stage model where a clinic wishes to maximize its profit by utilizing a single fixed resource to satisfy demand from two demand classes which differ with respect to no-show rates. The sequence of events is described graphically in Figure (4.2). We assume that appointment requests from each class arrive individually and independently with all requests from Class 2 patients arriving before the first request is received from Class 1 patients. In this setting, we assume Class 2 patients generally represent visits which are scheduled weeks into
the future, such as follow-up, diagnostic, or routine appointments, whereas Class 1 customers represent visits related to the patient’s immediate medical condition. We denote the total number of random arrivals from each class as $D_1$ and $D_2$ respectively with distribution functions $F_1()$ and $F_2()$ with means $\lambda_1$ and $\lambda_2$, respectively.

At the start of each booking period, before realizing the demand for that period, the clinic must decide how many appointment requests to accept during the period. If a request is rejected, we assume the patient leaves the system without cost to the clinic. The clinic gains a revenue $p$ for each scheduled patient who arrives for service. The clinic has a fixed capacity of $k$ units and incurs an overbooking cost $C(Z)$ where $Z$ is the number of scheduled patients that arrive for service. While overbooking often involves “working” a patient into the schedule, it also can result in staff overtime costs and provider productivity loss as the workday becomes longer than scheduled. Typically $C(Z)$ is defined such that $C(Z) = 0$ if $Z \leq k$ and convex for $Z > k$.

The probability that a patient from class $i = 1, 2$ actually arrives for service, once booked, is denoted by $\alpha_i$. We assume that this probability is independent across patients. We assume that $\alpha_2 \leq \alpha_1$ which is consistent with the relationship previously discussed that higher appointment delay leads to higher no-show rates. We assume that Class 1 customers attend with certainty, i.e. $\alpha_1 = 1$, mostly for tractability since our clinic data, as well as data from other studies, show that even patients who make their appointment less than one day before the appointment time still have some positive no-show probability. However, for actual clinic data, the probability of such “same-day no-shows” is fairly low (5-10%) compared to that of appointments scheduled far in advance (15-40%) (see Figure 4.1). Using these assumptions, let $Z_i(x_i)$ be a binomial random variable with parameters $x_i$ and $\alpha_i$ where $x_i$ denotes appointments booked from Class $i = 1, 2$ and $Z_i(x_i)$ denotes the random number of Class $i$ patients who attend the appointment.
Prior to realizing demand from Class $i = 1, 2$, the clinic decides to accept up to $b_i$ reservations in order to maximize its expected profits. Using the above assumptions we can formulate a stochastic dynamic program where the state variable is defined as the number of reservations booked from class 2 and class 1 respectively, $(x_2, x_1)$. Let $V_i(x_2, x_1)$ denote the optimal expected profit at the start of period $i$ with $x_2$ Class 2 reservations and $x_1$ Class 1 reservations already booked. No-show heterogeneity requires tracking the number of bookings from both classes individually, not just total bookings. Using this notation, we write the Bellman equations as follows:

$$V_2(0, 0) = \max_{b_2 \geq 0} E \left[ \{ V_1(\min(b_2, D_2), 0) \} \right]$$

(4.1)

$$V_1(x_2, 0) = \max_{b_1 \geq 0} E \left[ \{ V_0(x_2, \min(b_1, D_1)) \} \right]$$

(4.2)

$$V_0(x_2, x_1) = E [pZ_2(x_2) + pZ_1(x_1) - C(Z_2(x_2) + Z_1(x_1))]$$

(4.3)

Since all Class 2 appointments are made before any Class 1 requests are booked, booking limit decisions are made sequentially, not simultaneously. After observing the number of Class 2
appointments booked, the clinic determines the Class 1 booking limit, which may be a function
of Class 2 bookings. Let \( b_1(x_2) \) denote the Class 1 booking limit given \( x_2 \) patients have been
scheduled. This value will always be at least as high as \( b_1(b_2) \), the Class 1 limit when the
Class 2 limit is reached, so at least \( b_1(b_2) \) units will be held for Class 1 requests. Using revenue
management terms, we can think of \( b_1(b_2) \) as the Class 1 protection level. When no overbooking
is allowed, the protection level for Class 1 will equal the difference between capacity and Class
2 booking limit (Talluri and van Ryzin 2005). When overbooking is allowed, the sum of the
Class 2 booking limit and the Class 1 protection level may be above capacity, and the tradeoff
between booking limits need not be linear or one-to-one as the two classes have different no-show
rates. Consider an example where capacity is 24 but the clinic initially books up to 25 Class
2 requests and protects 6 slots for Class 1 requests, overbooking by 7. If the clinic receives 15
Class 2 appointment requests, it might now book up to 12 Class 1 patients, not 16, because
Class 1 patients are more likely to attend than Class 2. Thus the actual overbooking pad is 3.

We define \( C(Z_2(x_2) + Z_1(x_1)) = h(Z_2(x_2) + Z_1(x_1) - k)^+ \) where \( h \) is the per-patient over-
time cost. The linear overtime cost assumption is common in previous literature (Talluri and van
Ryzin 2005), although in practice the marginal cost of overtime may be increasing as patients
wait longer and providers work more overtime. Therefore, nonlinear cost structures are possible
and also fit the framework of the model. We assume capacity is soft, which means the clinic
can always complete its overtime work without spillover into the next period, when in practice
some customers promised an appointment today may be asked to return tomorrow. We assume
\( h > p \) to avoid the trivial case where every request is accepted since revenue-net-overtime cost
per patient would be non-negative. For convenience, let \( L(x_2, x_1) = E(Z_2(x_2) + Z_1(x_1) - k)^+ \)
denote the expected number of overtime patients, i.e. expected overtime work in number of
patients, and rewrite $V_0(x_2, x_1)$ as follows:

$$V_0(x_2, x_1) = p\alpha x_2 + px_1 - hL(x_2, x_1)$$  \hspace{1cm} (4.4)

There are several differences between our model and previous revenue management models. We assume the clinic is only reimbursed for service if the appointment is attended, whereas payment at the time of reservation is common in the literature (Talluri and van Ryzin 2005). The two models can be shown to be interchangeable. In reality, third-party payers make it difficult for clinics to collect payment in advance of service, thus justifying our assumption. We assume the clinic commits to its booking limit before demand is realized. If the optimal decision variable $b_i$ does not depend on the demand distribution $D_i$ as shown in traditional models, the timing is irrelevant.

Four assumptions of our model merit further discussion. These relate to service of appointments, patient preferences, multiple resources, and rejection costs. First, we assume the clinic incurs an overbooking cost only if the total number of attending patients exceeds the total capacity. We ignore service time variation, spillover between service periods, and in-clinic wait time, which are key features of models which focus on appointment scheduling. In reality a clinic with capacity of 3 patients per hour could have 5 patients show up one hour and only 1 patient the next. In this case, no waiting cost would be incurred in our model since attended appointments do not exceed capacity aggregated over the two-hour interval. While previous literature considers appointment scheduling on a micro level, e.g. (Muthuraman and Lawley 2008), we aggregate the costs to a daily level to consider an appointment allocation decision instead of assuming that scheduling of all classes is terminated simultaneously.
Second, we assume patients prefer a single appointment date and are indifferent to appointment time. In reality, patients may have preferences for certain times or a range of dates. In reality, patients who are rejected from their first preference might attempt to schedule for other dates and times within their preference range.

Third, our model considers a single resource when in reality more complex dynamics arise from managing multiple resources. If patients are unable to obtain an appointment with their regular provider, they may choose service from a different in-clinic provider, a competitor, or a costly emergency provider. Analytical models which study care teams or multiple resources have seen little development in outpatient scheduling literature, an exception being Gupta and Wang (2008). However, such network models have been developed for revenue management applications in other industries.

Fourth, we assume there is no cost for rejecting appointment requests. This is a common assumption in previous literature (Talluri and van Ryzin 2005). Rejecting a request may imply future congestion or penalties such as lost business or increased emergency room demand, though the clinic may not incur these costs directly. These rejection costs may also differ across classes, further impacting the allocation decision. A high rejection total may indicate a problem with overall capacity, which we assume to be fixed. We do not consider rejection penalties explicitly, but we model an opportunity cost for rejecting appointment requests, i.e. the potential lost revenue the clinic could earn if the rejected appointment is attended without increasing the overtime cost.
4.3.2 Model Analysis

Given that the clinic’s decisions are made sequentially, we proceed with a two-stage analysis in order to optimize the model using backward induction. In the first stage we wish to determine the optimal number of Class 1 patients the clinic is willing to book given \( x_2 \) Class 2 patients have already been booked.

**Proposition 4.1.** The optimal booking limit for Class 1 as a function of the number of Class 2 patients booked is determined as 
\[
b_1(x_2) = \max \{ x_1 \geq 0 : P(Z_2(x_2) > k - x_1) \leq \frac{p}{h} \}
\]

**Proof.** All proofs appear in the appendix.

An important property of the above result is that the optimal booking limit for Class 1 does not depend upon \( D_1 \). Even if \( D_1 \) is known before the booking decision is made, or if the clinic may alter its booking limit within the booking period, the clinic still accepts up to the same number of appointment requests from Class 1. It is also important to observe that as \( x_2 \) increases by one patient, \( b_1(x_2) \) either remains constant or decreases by one patient. Likewise, the total number of patients scheduled, from both classes, increases by one or remains constant.

If \( Z_2(x_2) \) is approximated using a continuous distribution with CDF given by \( F_{x_2}(.) \) then the optimality condition holds at equality:

\[
b_1(x_2) = k - F_{x_2}^{-1}\left(1 - \frac{p}{h}\right)
\]  \hspace{1cm} (4.5)

Before analyzing Stage 2, we discuss a model presented by Talluri and van Ryzin (2005) where customer classes differ by prices instead of no-show rates. They assume demand for each
class is realized before the booking decision is made and that prices are collected at time of booking. Given the homogeneous no-show rate $\alpha$, let $y$ represent total reservations on hand and $Z(y)$ be a binomial random variable with mean $y\alpha$ and variance $y\alpha(1-\alpha)$. Assuming linear overbooking costs, the Bellman equations are:

$$\hat{V}_2(0) = E\left[\max_{0 \leq u \leq D_2}\{p_2u + \hat{V}_1(u)\}\right]$$

$$\hat{V}_1(y) = E\left[\max_{0 \leq u \leq D_1}\{p_1u + \hat{V}_0(y+u)\}\right]$$

$$\hat{V}_0(y) = -hE\left[Z(y) - k\right]^+$$

Talluri and van Ryzin (2005) show the value function has non-increasing differences since the overbooking cost function is convex with respect to the Class 2 reservations. By definition, a set of random variables $\{X(\theta), \theta \in \Theta\}$ is stochastically increasing convex (SICX) if $E[\phi(X(\theta))]$ is increasing convex for all increasing convex functions $\phi$ (Shaked and Shanthikumar 2007). Using the definition of sample path convexity, which implies stochastic convexity, both Shaked and Shanthikumar (2007) and Talluri and van Ryzin (2005) show that a binomial random variable with mean $np$ and variance $np(1-p)$ is SICX. Thus, $\hat{V}_0$, the negative expected overbooking cost, is concave since $h(x-k)^+$ is a convex function of $x$. The concavity of $V_i(y)$ for $i \geq 1$ follows by induction from the properties that the sum of concave functions is concave and that the component-wise maximum of a concave function is concave.

Letting $\Delta\hat{V}_{j-1}(y) = \hat{V}_{j-1}(y) - \hat{V}_{j-1}(y-1)$, the optimal booking limits for the Talluri and
van Ryzin (2005) model are

\[ \hat{s}_j = \max \{ y \geq 0 : p_j \geq \Delta V_{j-1}(y) \} \quad (4.9) \]

Here \( \hat{s}_j \) represents the limit on total reservations on schedule in period \( j \) not just Class \( j \) reservations accepted in the period \( j \). Given \( x \) reservations on hand the optimal Class 1 and Class 2 booking limits for this model are:

\[ \hat{b}_1(x) = \max \left\{ y \geq 0 : P(Z(x+y+1) > k-1) \leq \frac{p_1}{h\alpha} \right\} = \hat{s}_1 - x \]

\[ \hat{b}_2 = \max \{ y \geq 0 : \]

\[ P \left( D_1 \geq \hat{b}_1(y-1) \right) \leq \frac{p_2}{p_1} - \frac{h\alpha}{p_1} \sum_{d=k-x+1}^{\hat{b}_1(x)} P(D_1 = d) P(Z(x+d-1) > k-1) \]

In the Talluri and van Ryzin (2005) model, the optimal Class \( i \) booking limit does not depend upon Class \( i \) demand due to concavity of the value function. The clinic books up to the same limit for a class regardless of how demand for that class arrives. In the next section, we show how the above results regarding concavity and booking limits which are independent of arrival process may not necessarily hold when no-show rates differ across demand classes.
4.3.3 Structural Properties

Returning to analysis of our model with heterogeneous no-show rates, presented in Equations (4.1), (4.2), and (4.3), we proceed to the analysis for Stage 2. At the start of Stage 2, the clinic must determine the optimal number of Class 2 patients to book given that up to $b_1(x_2)$ Class 1 patients will be booked in the following period. In this section we characterize some properties about the value function assuming $D_2 = d_2$ is constant. Under this assumption, we rewrite the value function as

$$V_2(0,0) = \max_{b_2 \geq 0} V_1(\min(b_2,d_2),0) = \max_{0 \leq x_2 \leq d_2} V_1(x_2,0)$$

**Proposition 4.2.** If the Class 1 booking limit is constant, i.e. $b_1(x_2) = b_1 \forall x_2$, the expected profit at the start of Stage 1, $V_1(x_2,0)$, is concave in the number of Class 2 reservations, $x_2$.

Clearly, it is unlikely that $b_1(x_2)$ is constant for all values of $x_2$; however, the above proposition does show that the marginal expected profit is decreasing anytime $b_1(x_2) = b_1(x_2 - 1) = b_1(x_2 - 2)$. Next, we use stochastic ordering to show an important structural result when the total number of bookings, denoted by $s \geq k$, is fixed.

**Proposition 4.3.** If the total number of appointments scheduled is constant, i.e. $b_1(x_2) = s - x_2$ for all $x_2$ and some fixed $s \geq k$, then $\mathbb{E}[Z_2(x_2) + \min(D_1,b_1(x_2)) - k]^+$, the expected overbooking cost, is increasing up to $x_2$ such that $P(D_1 \geq s - x_2 + 1) \leq \alpha_2$ and decreasing for $x_2$ such that $P(D_1 \geq s - x_2 + 1) \leq \alpha_2$.

For a fixed $s$, Proposition 3 does not directly state that the overbooking cost function is convex or concave. It does state that the function is first increasing and then decreasing.
Therefore, the function will not be convex over all values of $x_2$. The two propositions provide some characterization of $V_1(x_2, 0)$, but they do not guarantee concavity since $b_1(x_2)$, the booking limit for Class 1, depends upon $x_2$, the number of class 2 reservations made. To analyze the function further we need to derive the first differences and second differences explicitly.

Assuming $x_2 - 1$ Class 2 patients have already booked and the $x_2^{th}$ appointment request has been received, let $\Delta V_1(x_2, 0) = V_1(x_2, 0) - V_1(x_2 - 1, 0)$ be the marginal expected profit of accepting the $x_2^{th}$ Class 2 demand:

$$\Delta V_1(x_2, 0) = pE[Z_2(x_2)] + pE[\min(b_1(x_2), D_1)] - hE[Z_2(x_2) + \min(b_1(x_2), D_1) - k^+]$$

$$- pE[Z_2(x_2 - 1)] + pE[\min(b_1(x_2 - 1), D_1)]$$

$$- hE[Z_2(x_2 - 1) + \min(b_1(x_2 - 1), D_1) - k^+] .$$

Denote the marginal change in the expected number of overtime patients as

$$\Delta L(x_2, b_1(x_2)) = E[Z_2(x_2) + \min(b_1(x_2), D_1) - k^+] - E[Z_2(x_2 - 1) + \min(b_1(x_2 - 1), D_1) - k^+]$$

Using this notation we can write

$$\Delta V_1(x_2, 0) = p\alpha_2 + p(E[\min(b_1(x_2), D_1)] - E[\min(b_1(x_2 - 1), D_1)])$$

$$- h\Delta L(x_2, b_1(x_2)) .$$

The above expression for the marginal profit can be simplified for two cases corresponding
to when \( b_1(x_2) = b_1(x_2 - 1) \) and \( b_1(x_2) = b_1(x_2 - 1) - 1 \).

**Proposition 4.4.** If \( b_1(x_2) = b_1(x_2 - 1) \),

\[
\Delta V_1(x_2, 0) = po_2 - ho_2S(x_2, b_1(x_2)) - ho_2P(Z_2(x_2 - 1) > k - b_1(x_2 - 1) - 1)P(D_1 \geq b_1(x_2 - 1)).
\]

If \( b_1(x_2) = b_1(x_2 - 1) - 1 \),

\[
\Delta V_1(x_2, 0) = po_2 - ho_2S(x_2, b_1(x_2)) - P(D_1 \geq b_1(x_2 - 1))(p - h(1 - o_2)P(Z_2(x_2 - 1) > k - b_1(x_2 - 1))).
\]

where \( S(x_2, b_1(x_2)) = \sum_{d=k-x_2+1}^{b_1(x_2)-1} P(D_1 = d)P(Z_2(x_2 - 1) > k - d - 1) \)

Note that monotonicity of the marginal expected profit cannot be derived straightforwardly from the simplified expressions for the two cases. Even when looking at the second differences, it is difficult to guarantee the conditional profit function will be concave or even unimodal in \( x_2 \). Figure 4.3 shows the expected profit function given unlimited class 2 demand across three scenarios with \( k = 24 \) and \( p = 100 \). In two of the examples, when \( o_2 = 0.6 \), the conditional profit function is not unimodal; this is more difficult to see in the case where \( \lambda_1 = 18 \) because the modes are only two apart. When the function is multimodal, one cannot find the global optimal
solution using standard marginal analysis techniques, thereby making the problem more difficult to solve. Furthermore, in these two cases, the globally optimal value of \( x_2 \) is not the first local optimum. This makes finding the optimal solution even more difficult as the optimal policy will now depend upon the distribution of \( D_2 \). For the simple, two-stage model presented here, the optimal solution can still be found by completing an exhaustive search of all policies; however, a closed form expression is not possible. This search requires conditioning upon all values of Class 2 demand unless we can show the first local optimum is globally optimal. Even if the function is unimodal, the first order conditions do not provide a particularly simple expression for the optimal policy, though the search time to find the optimal policy is dramatically decreased.

### 4.3.4 Bounds on Optimal Solution

While it is difficult to find a closed form expression for the optimal class 2 booking limit, two simple approximations serve as lower and upper bounds respectively.
**Lower Bound** A lower bound can be derived by a variation on Littlewood’s rule (Littlewood 1972) which provides the optimal class 2 booking limit if no overbooking is allowed. In this case, let \( b_1(x_2) = k - x_2 \) and \( x_2 = \min(D_2, b_2) \) and write

\[
V_1(x_2, k - x_2) = po_2 x_2 + pE[\min(D_1, k - x_2)]. \tag{4.11}
\]

Using marginal analysis, we find a variation of Littlewood’s rule where \( p_2 = po_2 \) and \( p_1 = p \). With no overbooking, we accept the \( x_2^{th} \) patient if and only if

\[
\Delta V_1(x_2, k - x_2) = po_2 - pP(D_1 > k - x_2) \geq 0. \tag{4.12}
\]

We can guarantee a single optimum exists because \( P(D_1 > k - x_2) \) is increasing in \( x_2 \). The optimal policy if no overbooking is allowed is given by

\[
b_{2w}^{lw} = \max \{ x_2 : P(D_1 > k - x_2) \leq o_2 \}. \tag{4.13}
\]

The lower bound is therefore a special case of the of the original problem where Class 2 appointments attend with certainty but the clinic only receives the expected revenue \( po_2 < p \). In this case, \( b_1(x_2) = k - x_2 \) and the clinic never overbooks because \( h > p \). If Class 2 requests have a positive probability of failing to attend, the clinic books at least as many Class 2 requests as in the case when all show up with certainty, and thus \( b_{2w}^{lw} \) is a lower-bound for the optimal \( b_2 \).

Using this lower bound, we also gain insight into conditions for the optimality of a policy, such as same-day scheduling or pure-open access, which does not book class 2 requests, i.e. \( x_2 = 0 \). For such a policy to be superior to all policies, including those that overbook, it must
also be superior to all policies which do not overbook which implies $P(D_1 > k - 1) > \alpha_2$ or similarly $P(D_1 \geq k) > \alpha_2$. However, this condition does not guarantee that a pure open access policy will be optimal as later demonstrated numerically in Figure 5.

**Upper Bound** To derive an upper bound on Class 2 reservations, assume the clinic does not allow Class 1 patients to book, i.e. $b_1(x_2) = 0$. In this case, the marginal expected profit of accepting an additional request given $x_2 - 1$ have already been booked is:

$$
\Delta V_0(x_2, 0) - \Delta V_0(x_2 - 1, 0) = \alpha_2 p - h\alpha_2 P(Z_2(x_2 - 1) > k - 1).
$$

(4.14)

Since the expected marginal profit is decreasing in $x_2$, the clinic accepts Class 2 patients until the expected marginal profit is less than zero or equivalently:

$$
b^{ub}_2 = \max \left\{ x_2 : P(Z_2(x_2 - 1) > k - 1) \leq \frac{p}{h} \right\}.
$$

(4.15)

### 4.3.5 Sensitivity Analysis

Since we cannot derive a closed-form expression for the optimal solution, we perform sensitivity analysis using computational search methods. We examine changes in the optimal profit and booking limits relative to parameter values reflecting different clinical conditions, and we test the performance of the upper and lower bounds over these parameter values. We enumerate all feasible policies, and search over all policies where the clinic chooses the optimal Class 1 booking limit given a choice of the Class 2 booking limit. We assume Poisson distributions for the demand from each class with mean parameters $\lambda_1$ and $\lambda_2$ respectively. The base case
scenario for the sensitivity analysis uses parameter values: $k = 24; \alpha_2 = 0.75; \lambda_1 = 10; \lambda_2 = 20; p = 100; h = 150$.

In Figure 4.4, we see that optimal profits are increasing in the Class 2 attendance rate, the mean Class 1 demand, and the revenue-to-cost ratio. In Figure 4.a we see that the optimal expected profit increases as the Class 2 attendance rate increases, and in Figure 4.b we find that the optimal class 2 booking limit is generally decreasing in the Class 2 attendance rate with some minor jumps where it increases due to the discrete nature of the problem. This result agrees with our intuition that clinic profits should increase as both the no-show rate and resulting variability in service arrivals decrease. While the expected revenue from booking a Class 2 customer relative to a Class 1 customer increases with the Class 2 attendance rate, the clinic chooses to book fewer Class 2 patients as the effects of overbooking outweigh the trade-offs between class allocations. In Figure 4.c and 4.d we see that the optimal profit is increasing in $\lambda_1$ and the optimal booking limit is decreasing in $\lambda_1$. Since there is no cost for rejecting patient requests, clinic profits should also increase when mean demand increases because the clinic cannot do any worse by receiving more requests. Intuitively, the clinic should allocate fewer slots to Class 2 patients when the probability of an additional Class 1 customer is higher. In Figures 4.e we see that the optimal profit is increasing in the $p/h$ ratio as the expected marginal revenue of scheduling an additional patient increasingly outweighs the expected marginal cost. In Figure 4.f the optimal booking limit is increasing in the $p/h$ ratio as the clinic is willing to take on more risk.

Across all of the graphs in Figure 4 we see that for each individual parameter, when all other parameters are held constant, the gap between the upper and lower bound booking limit is either monotonically increasing or decreasing. In the case of sensitivity to the Class 2 attendance rate,
Figure 4.4: Optimal Policy and Profit Sensitivity Analysis

(a) Expected Profit vs. Class 2 Attendance Rate
(b) Class 2 Booking Limit vs. Class 2 Attendance Rate
(c) Expected Profit vs. Mean Class 1 Demand
(d) Class 2 Booking Limit vs. Mean Class 1 Demand
(e) Expected Profit vs. $ph = \text{Revenue-to-Cost Ratio}$
(f) Class 2 Booking Limit vs. $ph = \text{Revenue-to-Cost Ratio}$
the lower bound value is increasing as the upper bound value is decreasing. For the other two sensitivity parameters, one bound stays constant (upper bound for Class 1 demand and lower bound for $p/h$ ratio) while the other converges with the optimal value. Likewise, for each individual sensitivity parameter, the optimal policy tends to be closer to one bound at one end of the parameter range and closer to the other bound at the other end of the parameter range. When the Class 2 attendance rate is low, the optimal policy is closer to the upper bound; however, as the Class 2 attendance rate increases, both bounds converge toward the optimal solution with the lower bound showing a greater improvement. When the mean Class 1 demand, $\lambda_1$ is low, the optimal Class 2 booking limit is close to the upper bound and when the mean Class 1 demand is high, the optimal Class 2 booking limit is close to the lower bound. When the $\frac{p}{h}$ ratio is low, revenue per customer is low relative to cost, and the optimal Class 2 booking limit is close to the lower bound. When $\frac{p}{h}$ is high, the optimal Class 2 booking limit is close to the upper bound since the clinic has more incentive to risk overtime cost. Further sensitivity analysis in Section 4.3 evaluates changes in policy performance with respect to changes in the parameters.

### 4.4 Numerical Study

We build a numerical study to determine the economic value of joint overbooking and capacity control in a clinic setting, test simple approximations, and provide insights for clinic managers across a variety of parameter settings. In the first subsection, we describe the numerical study design which involves 198 scenarios from a wide range of model parameters and ten unique policies developed from our results, previous literature, and common practice. In the second subsection, we compare each policy’s expected profit performance by looking at percent dif-
ference from optimal, percent improvement over first-come-first-serve, relative rankings, and counts of optimal and near-optimal scenarios for each policy. In the third subsection, we discuss conditions where a pure Open Access policy achieves optimal or near-optimal expected profit. In the fourth subsection 4.4, we investigate differences in policy decision variables and other clinic performance measures. In the fifth subsection, we analyze the sensitivity of expected profits to model parameters for all policies to test the robustness of our results. Finally, we provide managerial insights on when to use certain approximations over others.

4.4.1 Experimental Design

We use a full-factorial design of the following parameters and consider 198 scenarios (out of 225 total) in which mean total demand is at least capacity, \( \lambda_1 + \lambda_2 \geq k \). We build our scenarios using the following set of parameters: \( k = 24 \), \( \alpha_2 \in \{0.6, 0.75, 0.9\} \), \( \lambda_i \in \{6, 12, 18, 24, 30\} \), \( p/h \in \{0.25, 0.5, 0.75\} \); \( p = 100 \). The base case refers to the additional scenario used in the earlier Section 4.3 with parameter values: \( k = 24 \); \( \alpha_2 = 0.75 \); \( \lambda_1 = 10 \); \( \lambda_2 = 20 \); \( p = 100 \); \( h = 150 \). Optimal base case expected profits for each policy are denoted by \( V_{opt}^policy \). We now describe the policies tested, beginning with those developed earlier in this work.

- Optimal (OPT): Policy is determined by global search of full policy enumeration. We compute the expected profit for each policy by conditioning over all values of Class 2 demand, Class 1 demand, and Class 2 no-shows. We reduce computation time by searching within upper and lower bounds presented in Section 3.4.

\[
- b_2^{opt} = \arg\max_{b_2} \{E_{D_2}[V_1(\min(D_2,b_2),b_1(\min(D_2,b_2)))]\}
\]

\[
- b_1(x_2) = \max \{x_1 : P(Z_2(x_2) > k - x_1) \leq \frac{p}{h}\}
\]

151
- Base case: $b_2^{opt} = 25; b_1 \left( b_2^{opt} \right) = 6; V_2^{opt} = 2189.89$

- Stopping Rule (STOP): Policy approximates the optimal Class 2 booking limit as the first local maximum of the conditional profit function, which is globally optimal if the function is unimodal and a lower bound otherwise. To derive policy values, we must evaluate the complex terms in the marginal profit equations.

  $- b_2^{stop} = \min \{ x_2 : \Delta V_1 (x_2 + 1, b_1 (x_2 + 1)) < 0 \}$

  $- b_1 (x_2) = \max \{ x_1 : P (Z_2 (x_2) > k - x_1) \leq \frac{p}{h} \}$

  $- Base case: b_2^{stop} = 25; b_1 \left( b_2^{stop} \right) = 6; V_2^{stop} = 2189.89$

- Marginal Revenue Update (MR): Policy approximates the Class 2 booking limit as the maximum value where the expected marginal revenue of accepting a Class 2 patient outweighs that of reserving the slot for a Class 1 patient. The policy essentially assumes the marginal overbooking cost is negligible across different Class 2 booking limits when using the respective optimal Class 1 booking limit. Since the optimal Class 1 booking limit for any number $x_2$ of Class 2 reservations must be at least $k - x_2$ by the earlier propositions, then the MR Class 2 booking limit must be at least as high as the lower bound in Section 3.4.

  $- b_2^{mr} = \max \{ x_2 : P (D_1 > b_1 (x_2)) \leq \alpha_2 \}$

  $- b_1 (x_2) = \max \{ x_1 : P (Z_2 (x_2) > k - x_1) \leq \frac{p}{h} \}$

  $- Base case: b_2^{mr} = 23; b_1 \left( b_2^{mr} \right) = 8; V_2^{mr} = 2188.74$

- Lower Bound (LB): The lower bound approximation for the Class 2 booking limit and respective optimal Class 1 booking limit is computed using expression (4.13).
- $b_2^{lb} = \max \{ x_2 : P(D_1 > k - x_2) \leq \alpha_2 \} = b_2^{lw}$
- $b_1(x_2) = \max \{ x_1 : P(Z_2(x_2) > k - x_1) \leq \frac{P}{\pi} \}$
- Base case: $b_2^{lb} = 16; b_1 (b_2^{lb}) = 13; V_2^{lb} = 2088.22$

- Upper Bound (UB): The upper bound approximation for the Class 2 booking limit and respective optimal Class 1 booking limit is computed using expression (4.15).

- $b_2^{ub} = \max \{ x_2 : P(Z_2(x_2 - 1) > k - 1) \leq \frac{P}{h} \}$
- $b_1(x_2) = \max \{ x_1 : P(Z_2(x_2) > k - x_1) \leq \frac{P}{\pi} \}$
- Base case: $b_2^{ub} = 33; b_1 (b_2^{ub}) = 0; V_2^{ub} = 2189.33$

The second group of policies for comparison includes solutions to traditional versions of the problem which impose stricter assumptions.

- Littlewood Rule (LW): The best policy which does not overbook, a straightforward application of Littlewood’s rule.

- $b_2^{lw} = \max \{ x_2 : P(D_1 > k - x_2) \leq \alpha_2 \}$
- $b_1(x_2) = k - x_2$
- Base case: $b_2^{lw} = 16; b_1 (b_2^{lw}) = 8; V_2^{lw} = 1944.16$

- Single Class (SC): Policy uses FCFS allocation and a single overbooking limit which assumes all patients have the same no-show rate equal to the weighted average of the mean demands and no-show rates of the two classes.

- $\alpha = \frac{(\lambda_1 \alpha_1 + \lambda_2 \alpha_2)}{\lambda_1 + \lambda_2}$
- $Z(x) \sim Bin(x, \alpha)$
\[
- b_2^{sc} = \max \left\{ x : P(Z(x-1) > k-1) \leq \frac{p}{h} \right\} \leq x_{ub}
\]
\[
- b_1(x_2) = b_2^{sc} - x_2
\]
- Base case: \( b_2^{sc} = 30; b_1(b_2^{sc}) = 0; V_2^{sc} = 2182.54 \)

- Revised Open Access (ROA): Policy sets Class 1 booking limit at capacity and chooses the optimal corresponding value for the Class 2 booking limit. This policy guarantees all same-day patients an appointment (up to capacity), while allowing some advance bookings to hedge against risk of low same-day demand.

\[
b_2^{roa} = \max \{ x_2 : p \alpha_2 - h \alpha_2 P(D_1 \geq k) \\
- h \alpha_2 \sum_{d=0}^{k-1} P(D_1 = d) P(Z_2(x_2 - 1) > k - d - 1) \geq 0 \}
\]

- \( b_1(x_2) = k \)
- Base case: \( b_2^{roa} = 21; b_1(b_2^{roa}) = 24; V_2^{roa} = 2147.31 \)

The final group of policies encompasses those which are commonly used in practice.

- Pure Open Access (POA): Policy books Class 1 to capacity, no advance booking.

- \( b_2^{oa} = 0 \)
- \( b_1(x_2) = k \)
- Base case: \( b_2^{oa} = 0; b_1(b_2^{oa}) = 24; V_2^{oa} = 999.99 \)

- First-Come-First Serve (FCFS): This policy satisfies all appointment requests, independent of patient class, until reaching capacity.
\[ b_{2}^{fcfs} = k \]

\[ b_{1}(x_2) = k - x_2 \]

- Base case: \( b_{2}^{fcfs} = 24; b_{1}(x_2) = 0; V_{2}^{fcfs} = 1878.75 \)

### 4.4.2 Policy Performance

In this section, we compare expected profits of the policies on the basis of expected difference from optimal, relative policy ranking, expected improvement over first-come-first-serve allocation, and counts of optimality or near-optimality for each policy.

Table 4.1 provides five summary statistics for the percentage difference from optimal expected profits for each of the nine policies across the 198 scenarios. The STOP policy performs very well but is the most difficult to compute. The median and mean percent errors are 0% and 0.09%, respectively, showing that the first local maximum of the conditional profit function tends to perform well compared to the global maximum. The MR policy also performs well with a median percent error of 0.04% and mean of 0.29%. The MR policy is much easier to compute than the STOP policy; however, MR may still require finding multiple values of \( b_{1}(x_2) \).

Given their computational simplicity, the LB and UB policies also perform well on average, but worst case scenarios are much worse than those of STOP and MR policies (19.60% and 26.87%, as opposed to 2.34% and 2.55% respectively). Across all scenarios, the average and standard deviation for the percent difference from optimal is lower for UB than is for LB (1.58% vs. 2.24% and 2.33% vs. 3.49%, respectively). One possible explanation for why UB outperforms LB on average can be seen in Figure 4.4, where the profit curve plateaus after reaching its maximum for some scenarios. Therefore, a booking limit well above optimal may perform better than a booking limit slightly below optimal. For example, when \( \lambda_1 = \lambda_2 = 18, \alpha_2 = 0.75, \) and
Table 4.1: Summary Statistics of Percent Difference from Optimal by policy

<table>
<thead>
<tr>
<th></th>
<th>STOP</th>
<th>MR</th>
<th>LB</th>
<th>UB</th>
<th>ROA</th>
<th>SC</th>
<th>LW</th>
<th>FCFS</th>
<th>POA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.09</td>
<td>0.29</td>
<td>2.24</td>
<td>1.58</td>
<td>3.86</td>
<td>3.95</td>
<td>4.38</td>
<td>13.63</td>
<td>23.00</td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>0.37</td>
<td>0.51</td>
<td>3.49</td>
<td>2.33</td>
<td>4.11</td>
<td>4.90</td>
<td>5.55</td>
<td>10.80</td>
<td>24.57</td>
</tr>
<tr>
<td>Minimum</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.000*</td>
<td>0</td>
<td>0.005</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Maximum</td>
<td>2.34</td>
<td>2.55</td>
<td>19.60</td>
<td>26.87</td>
<td>16.43</td>
<td>26.87</td>
<td>26.39</td>
<td>38.59</td>
<td>74.21</td>
</tr>
<tr>
<td>Median</td>
<td>0</td>
<td>0.04</td>
<td>0.55</td>
<td>2.01</td>
<td>2.76</td>
<td>2.01</td>
<td>2.34</td>
<td>9.61</td>
<td>16.96</td>
</tr>
</tbody>
</table>

* UB never exactly matches optimal policy. Actual minimum difference = 0.0000000227%

For $h = 200$, the optimal policy sets the Class 2 booking limit at 16, whereas the LB policy chooses 9 and the UB policy chooses 32. Though the LB limit is only 7 lower than optimal and the UB limit is 16 higher than optimal, the UB policy has a higher expected profit (2246.75 vs. 2204.80). Table 4.4 reveals that while UB outperforms LB on average, it is not optimal for any scenario, though expected profit is close in 34 scenarios, while LB is optimal in 51 scenarios. Also, the worst case performance for the UB policy is worse than that for the LB policy, as seen in Table 4.1.

The performances of SC and LW provide some insight into the value of overbooking and the value of considering multiple classes in patient booking. The best policy which does not overbook, LW, has expected profits 4.38% less than optimal on average across all scenarios. Similarly, SC, a policy which assumes a single no-show rate, has expected profits 3.95% less than optimal on average across all scenarios. Using MR, a relatively simple joint overbooking and capacity control policy, expected profits average only 0.44% less than optimal. Thus, using a joint overbooking and capacity control policy means that expected profits are 3.5-4% closer to optimal as opposed to either one alone. Table 4.2 displays how policies perform compared to a FCFS allocation policy. This table provides further insight into the value gained from using either overbooking or capacity control alone and the additional value from integrating the two practices. By incorporating optimal capacity control without overbooking (LW), expected
Table 4.2: Summary Statistics of Percent Increase over FCFS by policy

<table>
<thead>
<tr>
<th></th>
<th>%</th>
<th>OPT</th>
<th>STOP</th>
<th>MR</th>
<th>LB</th>
<th>UB</th>
<th>ROA</th>
<th>SC</th>
<th>LW</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum</td>
<td>0.005</td>
<td>0.005</td>
<td>0.005</td>
<td>0.005</td>
<td>0.002</td>
<td>-15.569</td>
<td>0</td>
<td>0.005</td>
<td></td>
</tr>
<tr>
<td>Maximum</td>
<td>62.832</td>
<td>62.832</td>
<td>62.832</td>
<td>62.832</td>
<td>57.312</td>
<td>62.832</td>
<td>56.662</td>
<td>62.832</td>
<td></td>
</tr>
</tbody>
</table>

Profits increase 12.3% above FCFS on average across all scenarios. By incorporating overbooking without capacity control, assuming a single weighted no-show rate, expected profits increase 12.6% above FCFS on average across all scenarios. All joint capacity control and overbooking policies outperform LW and SC with respect to increased expected profits over FCFS by 1.1-5.4% and 0.9% to 5.1%, respectively.

Summary statistics on the rank ordering of policies are compiled in Table 4.3. An interesting result is that neither SC nor FCFS is ever the best policy. In fact, the best rankings of these two policies are third and sixth, respectively. Another interesting result is that although the STOP policy is never more than 2.34% from optimal, its worst ranking is sixth. STOP tends to have a low ranking when the probability of Class 2 demand is significantly high, since the optimal solution is likely to be a local optimum other than the first. From further investigation, we determine that MR achieves its worst ranking of fourth in three scenarios where it is outperformed by STOP, UB, and SC respectively. These scenarios are characterized by moderately high mean Class 1 demand, low mean Class 2 demand, and high no-show rates. In these scenarios, the MR policy tends to underestimate the optimal Class 2 booking limit as it does not appropriately capture that the marginal expected overtime cost may actually be decreasing and a higher Class 2 booking limit may be preferable.

Table 4.4 presents the number of scenarios each policy matched the optimal policy and the
Table 4.3: Summary Statistics of Policy Rankings (1 highest expected profit, 9 lowest)

<table>
<thead>
<tr>
<th></th>
<th>STOP</th>
<th>MR</th>
<th>LB</th>
<th>UB</th>
<th>ROA</th>
<th>SC</th>
<th>LW</th>
<th>FCFS</th>
<th>POA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>1.268</td>
<td>1.970</td>
<td>3.495</td>
<td>5.404</td>
<td>5.783</td>
<td>4.490</td>
<td>8.136</td>
<td>6.975</td>
<td></td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>1.101</td>
<td>0.860</td>
<td>1.930</td>
<td>2.038</td>
<td>2.198</td>
<td>1.736</td>
<td>2.584</td>
<td>0.859</td>
<td>2.867</td>
</tr>
<tr>
<td>Best Rank</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>Worst Rank</td>
<td>6</td>
<td>4</td>
<td>3</td>
<td>8</td>
<td>8</td>
<td>7</td>
<td>9</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>Median Rank</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>6</td>
<td>8</td>
<td>9</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.4: Count of Scenarios (out of 198)

<table>
<thead>
<tr>
<th></th>
<th>STOP</th>
<th>MR</th>
<th>LB</th>
<th>UB</th>
<th>ROA</th>
<th>SC</th>
<th>LW</th>
<th>FCFS</th>
<th>POA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optimal</td>
<td>148</td>
<td>60</td>
<td>51</td>
<td>0</td>
<td>30</td>
<td>0</td>
<td>51</td>
<td>0</td>
<td>30</td>
</tr>
<tr>
<td>0.01% less than OPT</td>
<td>163</td>
<td>87</td>
<td>56</td>
<td>34</td>
<td>30</td>
<td>2</td>
<td>55</td>
<td>1</td>
<td>30</td>
</tr>
<tr>
<td>Best policy</td>
<td>185</td>
<td>73</td>
<td>64</td>
<td>3</td>
<td>30</td>
<td>0</td>
<td>64</td>
<td>0</td>
<td>30</td>
</tr>
</tbody>
</table>

number of scenarios each policy resulted in an expected profit less than 0.01% below optimal. The MR policy generalizes the LW policy and therefore dominates LW in performance. The simple approximations perform well, with LB less than 0.01% from optimal in 56 scenarios and UB less than 0.01% from optimal in 34 scenarios. Further investigation of results not captured in Table 4.3, reveals that both LB and UB are less than 0.01% from optimal in only two scenarios; meaning that in 88 distinct scenarios either LB or UB (or both) perform extremely well with expected profits less than only 0.01% from optimal. To provide further detail, for each scenario we let the Closer Bound Profit (CBP) denote the maximum expected profit achieved by either LB or UB. On average across all scenarios, we find that CBP is only 0.30% different from optimal with a median of 0.03%. Thus, using LB and UB with discretion can be an effective policy. Of the two policies, LB (UB) should be implemented for high (low) values of $\lambda_1$ and low (high) values of $p/h$. While performance improves for both policies as the Class 2 attendance rate increases, it improves more for LB, thus LB is also recommended for high values of $\alpha_2$. 

158
4.4.3 Optimality of Pure Open Access

As seen in Table 4.4, a pure Open Access policy can be optimal under certain conditions. We further analyze these results in this subsection. Open Access, in particular same-day scheduling, has received much recent attention in practice and in the literature, with little agreement regarding scenarios under which it will work best. In this section, we contribute to Open Access research by testing such conditions within our model.

In the numerical study, a pure Open Access policy is optimal whenever $P(D_1 \geq k) > \alpha_2$. When such a condition holds, the probability the clinic experiences enough Class 1 demand to fill capacity exceeds the attendance rate for Class 2 requests. The expected marginal revenue from giving the $k^{th}$ appointment slot to Class 1 is positive and all initial capacity (not including overbooking pad) should be allocated to Class 1. In the case where $\alpha_2 \geq p/h$, negative marginal expected profit is incurred for every additional patient overbooked in Class 2 beyond the $k$ slots allocated to Class 1. Not only should all $k$ slots of fixed capacity be allocated to Class 1, but allowing additional overbooking from Class 2 is suboptimal. In the numerical study, the Open Access policy appears to remain optimal when $p/h > \alpha_2$, so long as $P(D_1 \geq k) > \alpha_2$. This result, however, does not hold generally as in Figure 4.5 we see that for certain high values of $p/h$, even when $P(D_1 \geq k) > \alpha_2$, the optimal policy allows some Class 2 customers to book.

4.4.4 Other Performance Measures

We analyze other performance measures, besides expected profit, such as average booking limit values, average number of attended appointments, average number of booked appointments, and average number of rejected appointment requests.
Table 4.5 summarizes the class 2 booking limit (left column) and protection level (right column) across different policies. We omit FCFS and POA from the table as FCFS always sets the Class 2 Booking Limit at capacity and the protection level at 0, while POA always sets the Class 2 Booking Limit at 0 and the protection level at capacity. From the table we observe that on average across all scenarios the optimal policy will allow up to 56.2% of capacity to be booked by Class 2 patients and protect 58.4% of capacity for Class 1 patients. The additional 14.6% above capacity can be interpreted as the average percent of capacity the optimal policy would be willing to overbook if the Class 2 booking limit was reached, or equivalently, the maximum amount of overbooking allowed. By comparison, the best policy which does not overbook (LW) protects an average 65.8% of capacity for Class 1 patients and allows Class 2 patients to book up to 34.2% of capacity. This provides an interesting comparison; though the optimal policy is willing to book more total patients than LW through the use of overbooking, it protects a smaller percentage of its capacity for Class 1 patients on average. One explanation for this difference is that when a clinic does not overbook, it has fewer appointments to allocate and prefers to save more of those limited appointments for the preferred Class 1 patients, who have
Table 4.5: Class 2 Booking Limit (left) and Class 1 Protection Level (right) by Policy

<table>
<thead>
<tr>
<th></th>
<th>OPT</th>
<th>STOP</th>
<th>MR</th>
<th>LB</th>
<th>UB</th>
<th>ROA</th>
<th>SC</th>
<th>LW</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>13.5</td>
<td>14.0</td>
<td>13.0</td>
<td>14.3</td>
<td>8.2</td>
<td>17.6</td>
<td>32.7</td>
<td>0</td>
</tr>
<tr>
<td>SD</td>
<td>9.8</td>
<td>6.9</td>
<td>9.8</td>
<td>7.0</td>
<td>9.8</td>
<td>5.6</td>
<td>5.8</td>
<td>0</td>
</tr>
<tr>
<td>Min</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Max</td>
<td>37</td>
<td>24</td>
<td>37</td>
<td>24</td>
<td>35</td>
<td>24</td>
<td>43</td>
<td>0</td>
</tr>
<tr>
<td>Med.</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>14</td>
<td>11</td>
<td>17</td>
<td>7</td>
<td>18</td>
</tr>
</tbody>
</table>

Another important finding from Table 5 is that the best performing policies tend to be more flexible, demonstrating higher standard deviations and ranges across scenarios. The best performing policies are more sensitive to changes in the parameter values across scenarios because they dynamically incorporate the no-show rate ($\alpha_2$), the Class 1 demand parameter ($\lambda_1$), and the revenue ($p$) and overtime cost ($h$) into the computation of booking limits. The worst performing policies have static Class 2 booking limits.

Table 4.6 displays additional performance measures, including the expected number of bookings across classes, the expected number of attended appointments, and the expected number of rejected appointment requests across classes. Table 4.6 summarizes actual bookings, attended, and rejections whereas Table 5 summarizes what the clinic is willing to book before any requests are received. Tables 4.5 and 4.6 show that for Class 2, the booking limit, average attended, and average booked are slightly lower for MR than both OPT and STOP on average. The average expected total number of bookings, attended, and rejections is similar across these three policies. The OPT, STOP, MR, UB, and ROA policies all serve an average of about 23 total patients (95.8% of nominal capacity) with UB and SC serving an average of about 22. However, some of these policies strongly outperform others in terms of expected profits due to variation in overtime costs. For instance, the ROA policy does little to control variation in Class 1 demand.
by always protecting its full capacity for Class 1 patients. While ROA sees an average of 23.0 patients, high variation in Class 1 bookings leads to overtime and lower expected profits than LB which only sees an average of 22.1 appointments. Similarly, UB books many Class 2 patients so the variation in the number that will attend is higher than with policies which book fewer class 2 patients.

The policy with the highest average expected number of bookings is the UB policy. This policy gives preference to Class 2 patients, who have a lower attendance rate. On average, the UB policy books 18.8 Class 2 patients, 14.0 of which attend, and rejects 10.9 total requests. POA and ROA have the highest average expected Class 1 bookings (17.4). The policy with the lowest average expected total bookings (highest rejections) is POA because it does not allow any Class 2 patients to book. After POA, the next lowest (highest rejections) are the LW and FCFS policies which, in addition to POA, do not allow overbooking. POA and LW reject mainly Class 2 patients while FCFS and SC primarily reject Class 1 patients because they do not limit Class 2 bookings.

Table 4.6 also provides insight into the costs of rejecting appointments. The range of the total number of rejections across all policies except for pure open access is relatively small (10.9, 15.5), though the rejections differ dramatically for the two classes. We can infer from these results that incorporating a common rejection cost for both classes will not greatly affect the relative performance of the policies, although the policy values may change. If rejection costs differ among classes, this additional heterogeneity will have further impact on the allocation decision.
Table 4.6: Alternative Performance Measures by Policy

<table>
<thead>
<tr>
<th>Category</th>
<th>OPT</th>
<th>STOP</th>
<th>MR</th>
<th>LB</th>
<th>UB</th>
<th>ROA</th>
<th>SC</th>
<th>LW</th>
<th>FCFS</th>
<th>POA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class 2 Book</td>
<td>11.1</td>
<td>10.8</td>
<td>9.5</td>
<td>7.7</td>
<td>18.8</td>
<td>7.6</td>
<td>18.4</td>
<td>7.7</td>
<td>17.4</td>
<td>0.0</td>
</tr>
<tr>
<td>Class 2 Attend</td>
<td>8.3</td>
<td>8.0</td>
<td>7.1</td>
<td>5.9</td>
<td>14.0</td>
<td>5.6</td>
<td>13.8</td>
<td>5.9</td>
<td>13.1</td>
<td>0.0</td>
</tr>
<tr>
<td>Class 1 Book</td>
<td>14.7</td>
<td>14.9</td>
<td>15.7</td>
<td>16.2</td>
<td>9.1</td>
<td>17.4</td>
<td>8.1</td>
<td>15.5</td>
<td>6.2</td>
<td>17.4</td>
</tr>
<tr>
<td>Total Book</td>
<td>25.8</td>
<td>25.7</td>
<td>25.2</td>
<td>23.9</td>
<td>27.9</td>
<td>25.0</td>
<td>26.5</td>
<td>23.2</td>
<td>23.7</td>
<td>17.4</td>
</tr>
<tr>
<td>Total Attend</td>
<td>23.0</td>
<td>23.0</td>
<td>22.8</td>
<td>22.1</td>
<td>23.1</td>
<td>23.0</td>
<td>21.9</td>
<td>21.4</td>
<td>19.3</td>
<td>17.4</td>
</tr>
<tr>
<td>Class 2 Reject</td>
<td>8.3</td>
<td>8.6</td>
<td>9.9</td>
<td>11.7</td>
<td>0.6</td>
<td>11.8</td>
<td>0.9</td>
<td>11.7</td>
<td>1.9</td>
<td>19.4</td>
</tr>
<tr>
<td>Class 1 Reject</td>
<td>4.6</td>
<td>4.4</td>
<td>3.7</td>
<td>3.1</td>
<td>10.3</td>
<td>1.9</td>
<td>11.3</td>
<td>3.9</td>
<td>13.1</td>
<td>1.9</td>
</tr>
<tr>
<td>Total Reject</td>
<td>12.9</td>
<td>13.0</td>
<td>13.5</td>
<td>14.8</td>
<td>10.9</td>
<td>13.7</td>
<td>12.2</td>
<td>15.5</td>
<td>15.1</td>
<td>21.3</td>
</tr>
</tbody>
</table>

*Average Expectation Over 198 Scenarios. Note Class 1 Book = Class 1 Attend because $\alpha_1 = 1$.

4.4.5 Policy Sensitivity Analysis

In this subsection, we analyze the sensitivity of the expected profit of the policies to changes in parameters. We are interested in the consistency of our revenue-based model results with those of authors who have used cost-based models, in order to establish generality across different clinic objectives and situations.

Tables 4.7 provides sensitivity analysis of how the expected profit changes with respect to the Class 2 attendance rate as measured by the average percent different from optimal across scenarios. All of the policies except for ROA and POA show an improvement as the attendance rate increases. This agrees with intuition that if advanced scheduling is allowed, high no-show rates imply higher risk for overtime costs. The lower bound policies, such as LB and LW, show a greater relative improvement than the upper bound policies, such as UB and SC, with respect to the Class 2 attendance rate. LB profits are less than 0.3% from optimal on average when $\alpha_2 = 0.9$. Consistent with results found by other authors (Qu et al. 2007, LaGanga and Lawrence 2007, Robinson and Chen 2010), the performance of POA and the performance of ROA improve as the no-show rate increases. Across different values of mean demand and cost parameters, POA performs much worse than simple alternatives even when the attendance rate
Table 4.7: Avg. Percent Difference from Optimal by Class 2 Attendance Rate

<table>
<thead>
<tr>
<th></th>
<th>%</th>
<th>STOP</th>
<th>MR</th>
<th>LB</th>
<th>UB</th>
<th>ROA</th>
<th>SC</th>
<th>LW</th>
<th>FCFS</th>
<th>POA</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_2 = 0.6$</td>
<td>0.219</td>
<td>0.579</td>
<td>4.471</td>
<td>1.913</td>
<td>2.777</td>
<td>6.090</td>
<td>7.244</td>
<td>23.120</td>
<td>21.563</td>
<td></td>
</tr>
<tr>
<td>$\alpha_2 = 0.75$</td>
<td>0.045</td>
<td>0.233</td>
<td>1.949</td>
<td>1.738</td>
<td>3.721</td>
<td>4.057</td>
<td>4.551</td>
<td>13.483</td>
<td>22.984</td>
<td></td>
</tr>
<tr>
<td>$\alpha_2 = 0.9$</td>
<td>0.002</td>
<td>0.054</td>
<td>0.285</td>
<td>1.096</td>
<td>5.089</td>
<td>1.713</td>
<td>1.332</td>
<td>4.272</td>
<td>24.461</td>
<td></td>
</tr>
</tbody>
</table>

is quite low. Even FCFS strongly outperforms POA and slightly outperforms ROA, when the attendance rate is high.

Table 4.8 summarizes how the mean percent difference from optimal expected profit changes with respect to changes in mean Class 1 demand. As earlier results indicate, lower bound policies (LB, LW) show improved performance as mean Class 1 demand increases, while upper bound policies (UB, SC) show decreased performance. This result is illustrated in 4.5 where the optimal Class 2 booking limit is close to the upper bound when mean Class 1 demand is low and close to the lower bound when it is high. The MR results are interesting in that performance worsens as mean Class 1 demand increases; however when mean Class 1 demand is very high, the policy performs extremely well with an average of 0% difference from optimal. POA shows dramatic improvement in performance as mean Class 1 demand increases. Alternatively, ROA results are mixed as the average difference first increases, then decreases. This is likely due to the fact that at first demand is small enough that the chance of booking up to capacity is very small. As demand increases, the risk of booking some Class 2 patients and booking up to capacity in Class 1 patients increases. Once the probability of booking up to capacity in Class 1 patients is high enough, the ROA policy becomes a pure open access policy and does not allow any advance bookings. These results support a finding from Qu et al. (2007) that nearly all appointments will be held open if Class 1 demand is significantly higher than capacity. Other authors, such as Liu et al. (2010) and Green and Savin (2008) find that high demand leads to poor performance in open access because they explicitly model the impact of denying patients
Table 4.8: Avg. Percent Difference from Optimal by Mean Class 1 Demand

<table>
<thead>
<tr>
<th></th>
<th>STOP</th>
<th>MR</th>
<th>LB</th>
<th>UB</th>
<th>ROA</th>
<th>SC</th>
<th>LW</th>
<th>FCFS</th>
<th>POA</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\lambda_1 = 6)</td>
<td>0.000</td>
<td>0.012</td>
<td>5.242</td>
<td>0.117</td>
<td>2.381</td>
<td>0.451</td>
<td>10.386</td>
<td>12.704</td>
<td>70.511</td>
</tr>
<tr>
<td>(\lambda_1 = 12)</td>
<td>0.002</td>
<td>0.187</td>
<td>3.990</td>
<td>0.328</td>
<td>5.617</td>
<td>1.557</td>
<td>8.241</td>
<td>13.166</td>
<td>44.116</td>
</tr>
<tr>
<td>(\lambda_1 = 18)</td>
<td>0.005</td>
<td>0.380</td>
<td>2.256</td>
<td>0.696</td>
<td>7.893</td>
<td>2.816</td>
<td>4.829</td>
<td>12.328</td>
<td>19.760</td>
</tr>
<tr>
<td>(\lambda_1 = 24)</td>
<td>0.092</td>
<td>0.733</td>
<td>1.241</td>
<td>1.688</td>
<td>3.159</td>
<td>4.877</td>
<td>1.598</td>
<td>13.604</td>
<td>3.921</td>
</tr>
<tr>
<td>(\lambda_1 = 30)</td>
<td>0.292</td>
<td>0.000</td>
<td>0.000</td>
<td>4.246</td>
<td>0.021</td>
<td>8.185</td>
<td>0.000</td>
<td>15.862</td>
<td>0.021</td>
</tr>
</tbody>
</table>

Table 4.9: Avg. Percent Difference from Optimal by Mean Class 2 Demand

<table>
<thead>
<tr>
<th></th>
<th>STOP</th>
<th>MR</th>
<th>LB</th>
<th>UB</th>
<th>ROA</th>
<th>SC</th>
<th>LW</th>
<th>FCFS</th>
<th>POA</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\lambda_2 = 6)</td>
<td>0.114</td>
<td>0.154</td>
<td>0.393</td>
<td>0.679</td>
<td>2.433</td>
<td>1.292</td>
<td>0.997</td>
<td>3.327</td>
<td>6.120</td>
</tr>
<tr>
<td>(\lambda_2 = 12)</td>
<td>0.097</td>
<td>0.245</td>
<td>1.044</td>
<td>1.038</td>
<td>3.512</td>
<td>2.174</td>
<td>2.569</td>
<td>7.934</td>
<td>15.877</td>
</tr>
<tr>
<td>(\lambda_2 = 18)</td>
<td>0.080</td>
<td>0.281</td>
<td>1.948</td>
<td>1.343</td>
<td>3.787</td>
<td>3.297</td>
<td>4.253</td>
<td>12.838</td>
<td>27.484</td>
</tr>
<tr>
<td>(\lambda_2 = 24)</td>
<td>0.082</td>
<td>0.337</td>
<td>3.099</td>
<td>1.932</td>
<td>4.347</td>
<td>5.157</td>
<td>5.883</td>
<td>18.258</td>
<td>28.522</td>
</tr>
<tr>
<td>(\lambda_2 = 30)</td>
<td>0.082</td>
<td>0.364</td>
<td>3.717</td>
<td>2.450</td>
<td>4.591</td>
<td>6.426</td>
<td>6.512</td>
<td>20.511</td>
<td>28.832</td>
</tr>
</tbody>
</table>

Sensitivity analysis for the effect of mean Class 2 demand on percent difference from optimal expected profit is shown in Table 4.9. Since Class 2 patients have a higher no-show rate, the percent difference from optimal increases on average for all policies except for STOP. The performance for STOP improves slightly because the likelihood that the first local maximum will be the global maximum is higher as \(\lambda_2\), the mean demand for Class 2, increases. When \(\lambda_2\) is extremely low, LB outperforms UB on average; however, as \(\lambda_2\) increases, UB outperforms LB on average. Likewise, \(\lambda_2\), the LW policy outperforms the SC policy on average, meaning that if the clinic can implement only capacity control or overbooking, it prefers capacity control. As \(\lambda_2\) increases, the SC policy outperforms the LW policy on average and the clinic would prefer single class overbooking to capacity control with no overbooking. Since POA does not allow any Class 2 patients to book, its performance clearly deteriorates as \(\lambda_2\) increases, because it incurs a higher opportunity cost by not allowing Class 2 patients.

Table 4.10 provides results regarding how mean performance changes with respect to the
Table 4.10: Avg. Percent Difference from Optimal by \( p/h \), \( (p = 100) \)

<table>
<thead>
<tr>
<th></th>
<th>STOP</th>
<th>MR</th>
<th>LB</th>
<th>UB</th>
<th>ROA</th>
<th>SC</th>
<th>LW</th>
<th>FCFS</th>
<th>POA</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p/h = 0.25 )</td>
<td>0.057</td>
<td>0.181</td>
<td>1.450</td>
<td>2.706</td>
<td>5.950</td>
<td>5.680</td>
<td>3.030</td>
<td>12.415</td>
<td>22.025</td>
</tr>
<tr>
<td>( p/h = 0.5 )</td>
<td>0.175</td>
<td>0.322</td>
<td>2.284</td>
<td>1.467</td>
<td>3.720</td>
<td>3.858</td>
<td>4.425</td>
<td>13.671</td>
<td>23.026</td>
</tr>
<tr>
<td>( p/h = 0.75 )</td>
<td>0.034</td>
<td>0.363</td>
<td>2.972</td>
<td>0.574</td>
<td>1.917</td>
<td>2.322</td>
<td>5.671</td>
<td>14.789</td>
<td>23.957</td>
</tr>
</tbody>
</table>

ratio of revenue and cost parameters. As the overtime cost, \( h \), increases, the lower bound policies, LB and LW, perform better, where the opposite is true for the upper bound policies, UB and SC. This result is expected because the upper bound policies are exposing the clinic to more overtime risk and perform better when the overtime cost parameter is lower. This result provides insight into the sensitivity analyses in earlier sections and Figure 4.5. The performance of MR policy improves as \( h \) decreases. Likewise, the performance of POA increases slightly as \( h \) increases. POA does not overbook, so changes in expected profits are not due to changes in \( h \), given the same values for other parameters. The improvement in performance is more an indication of how the optimal policy is changing and the decrease in the opportunity cost for POA of not booking Class 2. Alternatively, the performance of ROA decreases as \( h \) increases because ROA is a joint overbooking and capacity control policy which allows some overbooking unless POA is optimal. Both policies are similar in that they offer all Class 1 patients an appointment (up to capacity), so we might expect similar behavior with regard to the changes in \( h \). It is also interesting because the result agrees with Robinson and Chen (2010) that open access performs better when there are small overtime surcharges.

### 4.5 Conclusion

We analyze joint overbooking and capacity control decisions in the presence of class-dependent no-shows. Assuming immediate appointment requests attend with certainty, we derive a simple
expression for the optimal booking limit for immediate requests as a function of previously booked appointments which may no-show. Due to no-show heterogeneity, the profit function may not be unimodal in the number of advance appointments accepted, making it necessary to use computationally intensive search methods to derive the optimal advance booking limit. We develop upper and lower bounds which greatly reduce the size of the search space. We perform sensitivity analysis to investigate how model parameters affect the optimal expected profit and policy. We compare ten policies developed from this paper, previous models from the literature, and popular outpatient appointment scheduling methods. The policies we develop perform extremely well compared to the optimal.

Our results show that expected profits are improved by allocating total capacity and overbooking to distinct demand classes with different no-show rates, even with constant revenue per patient. Compared to a first-come-first-serve (FCFS) allocation, expected profits increase 12.3% on average, across a numerical study of 198 scenarios, when incorporating capacity control with no overbooking. Expected profits increase 12.6% using a single overbooking limit and maintaining a FCFS allocation. When jointly incorporating capacity control and overbooking, the optimal policy improves expected profits 17.8% over FCFS. On average, expected profits for the lower and upper bounds perform within 2.24% and 1.58% of optimal, respectively. We also develop a Marginal Revenue (MR) policy which quickly and accurately approximates the optimal solution with mean and median percent errors less than 0.3%.

The implications of our results demonstrate how appointment allocation and overbooking decisions should consider differences in patient attendance rates, forecasts of immediate appointment requests, and revenues relative to overbooking costs. On average, across our numerical study the optimal policy protects 58.3% of its appointment capacity for immediate requests with
a range of 12.5% to 100%. In 147 of 198 scenarios, the optimal policy allows future appointments to overbook. On average, the optimal policy sets its advance booking limit at 56.19% of capacity with an average overbooking pad of 14.6% of capacity. On average, 25.8 total appointments (107.5% of capacity) are booked (11.1 advance and 14.7 immediate). When the advanced booking limit is not reached, the clinic increases the number of immediate requests it is willing to accept, but at less than a one-to-one ratio because attendance rates are higher for immediate requests. The optimal policy generally books more advance appointments as revenues increase relative to costs and fewer advance appointments as advance attendance rate or demand for immediate appointments increases. Given the variability of optimal policy values between scenarios, we show the value of using bounds and approximations provided in this work to determine the best allocation for the setting, instead of naive FCFS allocations or crude guidelines from practice - e.g. 25-35% future appointments and 75% immediate appointments (Qu et al. (2007), Green et al. (2007)).

A pure open access policy (POA), which reserves 100% of appointment capacity for immediate requests, does not perform well on average with expected profits 23.0% below optimal. POA can be optimal if the advance appointment attendance rate is small relative to the immediate demand distribution and the marginal overtime cost is not too far below the marginal revenue. We develop policies, such as MR policy, which generalize pure open access and outperform it under more general assumptions.

Our sensitivity analysis shows that clinics can achieve higher expected profits by making efforts to impact important model parameters. Optimal expected profits increase with the appointment attendance rate, the mean demand for immediate requests, and the $p/h$ ratio. Clinics might attempt to increase profits, for instance, by lowering no-show rates via appoint-
ment reminders, or by increasing demand for immediate appointments by encouraging patients to schedule closer to the desired appointment date.

Extensions to this work could incorporate features such as cancellations, rejection costs, multiple booking attempts, and dynamic models. Cancellations have been incorporated in previous revenue management models assuming static no-show rates and memory-less distribution of time until cancellation. If heterogeneity in attendance rates is partly due to cancellations, intuition from our model suggests clinics would increase booking limits, since recourse is available for some appointment waste. Often no explicit rejection costs are incurred in reality, but these parameters could model the effects of competition or increased appointment delay for patients. Future work could also assume rejected requests make additional booking attempts at a later stage or appointment date. Our model can also be extended into a dynamic model to allow for simultaneous booking of different patient classes, such as new or return patients. Empirical studies have investigated patient attributes that determine probability of a no-show, but clinics may find it ethically inappropriate to use personal attributes for appointment allocation. Additionally, future research could extend our model to more than two booking stages, but no-show heterogeneity requires tracking reservations from each class; leading to exponential growth in the state space.

As clinic management technology develops, more clinics gain the ability to implement complex appointment allocation policies. While recent work in operations sciences has developed effective policies, there is still opportunity for improvement translating research into practice. Managers and researchers should continue to collaborate on developing effective policies and methods for using real-time data to drive decisions.
Appendix A

Competition

A.1 Proofs of Key Results

A.1.1 Proof of Proposition 2.1

Proof. We write the monopolist’s optimization problem as

\[
\max_{x \geq 0, \mu \geq 0} \pi = r \lambda - qx - c\mu \quad \text{(A.1)}
\]

\[
\beta x - \frac{h}{\mu - \delta x - \lambda} - z = 0 \quad \text{(A.2)}
\]

\[
\lambda \leq \Lambda \quad \text{(A.3)}
\]

\[
\lambda \geq 0 \quad \text{(A.4)}
\]

\[
\lambda < \mu \quad \text{(A.5)}
\]

Solving the first constraint, \( \lambda = \mu - \delta x - \frac{h}{\beta x - z} \). Substituting back into the profit function, we write the KKT Lagrangian as follows

\[
L = r \left( \mu - \delta x - \frac{h}{\beta x - z} \right) - qx - c\mu
- \eta_1 \left( \mu - \delta x - \frac{h}{\beta x - z} - \Lambda \right) + \eta_2 \left( \mu - \delta x - \frac{h}{\beta x - z} \right) \quad \text{(A.6)}
\]

We write the stability conditions as

\[
\frac{\partial L}{\partial x} = (r - \eta_1 + \eta_2) \left( \frac{h\beta}{(\beta x - z)^2} - \delta \right) - q = 0 \quad \text{(A.7)}
\]

\[
\frac{\partial L}{\partial \mu} = r - c - (\eta_1 - \eta_2) = 0 \quad \text{(A.8)}
\]

And the complimentary slackness condition
\[
\eta_1 \left( \mu - \delta x - \frac{h}{\beta x - z} - \Lambda \right) = 0 \tag{A.9}
\]

\[-\eta_2 \left( \mu - \delta x - \frac{h}{\beta x - z} \right) = 0 \tag{A.10}\]

From Equation (A.8) and non-negativity conditions, \( \eta_i \geq 0 \quad \forall i = 1, 2 \), we see that the profit function is linear in the capacity decision, \( \mu \), for fixed quality effort decision, \( x \). Therefore, if \( r < c \), then \( \eta_2 > 0 \) which implies that \( \lambda = 0 \), i.e the provider does not enter the market. If \( r > c \), then \( \eta_1 > 0 \) which implies \( \lambda = \Lambda \). Assuming \( r > c \) implies \( \lambda > 0 \). Therefore, for any \( x \), \( \mu - \frac{h}{\beta x - z} = \Lambda \). Solving Equation (A.7) for \( x \), we obtain

\[
x = \frac{z}{\beta} + \sqrt{\frac{ch}{\beta (q + c\delta)}} \tag{A.11}
\]

And the closed-form expression for capacity is obtained by substituting back into \( \mu - \delta x - \frac{h}{\beta x} = \Lambda \) to get

\[
\mu = \Lambda + \frac{\delta z}{\beta} + (2c\delta + q) \sqrt{\frac{h}{c\beta (q + c\delta)}} \tag{A.12}
\]

\[\square\]

### A.1.2 Proof of Theorem 2.1

**Proof.** The proof for the first case is trivial. If the market potential is not high enough for either provider to make profit capturing the full market potential then neither provider will enter.

Similarly, in the second case, if the market potential is not high enough for Provider 2 to enter, then Provider 1 acts as monopolist.

In the third case, Provider 2 can enter as a monopolist, but there is not enough market potential for both firms to break-even so both firms cannot operate in the market.

In the fourth case, there is enough market potential for both firms to enter the market. In this case, if Provider \( j = 1, 2 \) sets quality and capacity such that it captures \( \lambda_j \geq \Lambda_j \) then the best response for Provider \( i \neq j \) is to capture the remaining market potential as if it were a monopolist. The result follows from Proposition 2.1. \[\square\]

### A.1.3 Proof of Proposition 2.2

**Proof.** We write the Lagrangian as follows

\[
\mathcal{L} = r \lambda - c \mu - qx - k + \eta_1 \left( \beta x - \frac{h \lambda}{\mu - \delta x} - z \right) - \eta_2 (\lambda - \Lambda) \tag{A.13}
\]
The stability conditions are given by

\[-q + \eta_1 \left( \beta - \frac{h \lambda \delta}{(\mu - \delta x)^2} \right) = 0 \quad (A.14)\]

\[r - \frac{\eta_1 h}{\mu \Delta x} - \eta_2 = 0 \quad (A.15)\]

\[-c + \frac{\eta_1 h \lambda}{(\mu - \delta x)^2} = 0 \quad (A.16)\]

The complimentary slackness conditions are given by

\[\eta_1 \left( \beta x - \frac{h \lambda}{\mu \Delta x} - z \right) = 0 \quad (A.17)\]

\[\eta_2 (\lambda - \Lambda) = 0 \quad (A.18)\]

Solving, we eliminate the cases of $[\eta_1 = 0, \eta_2 = 0]$ and $[\eta_1 = 0, \eta_2 > 0]$ as infeasible, i.e. consumers will be utility neutral with the outside option as shown in previous literature. The interior solution, $[\eta_1 > 0, \eta_2 > 0]$, provides a critical point solution which leads to negative profits; therefore the firm can do strictly better by not entering the market. The only non-trivial solution is for the firm to capture the full market with consumers utility neutral to the outside option, $[\eta_1 > 0, \eta_2 = 0]$. Solving that case provides the result. \hfill \Box

### A.1.4 Proof of Theorem 2.2

**Proof.** Follows from Proposition 2.2 and structure of Theorem 2.1. \hfill \Box

### A.1.5 Proof of Proposition 2.3

**Proof.** We write the Lagrangian as follows

\[\mathcal{L} = r \lambda - c \mu - q x - k + \eta_1 (\beta x - z) - \eta_2 (\lambda - \Lambda) \quad (A.19)\]

The stability conditions are given by

\[-q + \eta_1 \left( \beta - \frac{h \lambda}{(\mu - \frac{\lambda}{\gamma x})^2 \gamma x^2} \right) = 0 \quad (A.20)\]

\[r - \frac{\eta_1 h}{\left( \mu - \frac{\lambda}{\gamma x} \right)^2 \gamma x} - \eta_2 = 0 \quad (A.21)\]
\[ -c + \frac{\eta_1 h}{(\mu - \lambda)^2} = 0 \] (A.22)

The complimentary slackness conditions are given by

\[ \eta_1 \left( \beta x - \frac{h}{\mu - \lambda} - z \right) = 0 \] (A.23)

\[ \eta_2 (\lambda - \Lambda) = 0 \] (A.24)

Solving, we eliminate the cases of \([\eta_1 = 0, \eta_2 = 0]\) and \([\eta_1 = 0, \eta_2 > 0]\) as infeasible. The interior solution, \([\eta_1 > 0, \eta_2 > 0]\), provides a critical point solution which leads to negative profits; therefore the firm can do strictly better by not entering the market. The only non-trivial solution is for the firm to capture the full market with consumers utility neutral to the outside option, \([\eta_1 > 0, \eta_2 = 0]\). Solving that case provides the result. \(\square\)

### A.1.6 Proof of Theorem 2.3

**Proof.** Follows from Proposition 2.3 and structure of Theorem 2.1. \(\square\)

### A.1.7 Proof of Proposition 2.4

**Proof.** We write the Lagrangian as follows

\[ \mathcal{L} = (r - qx) \lambda - c \mu - k + \eta_1 \left( \beta x - \frac{h}{\mu - \delta x - \lambda} - z \right) - \eta_2 (\lambda - \Lambda) \] (A.25)

The stability conditions are given by

\[ -\lambda q + \eta_1 \left( \beta - \frac{h \delta}{(\mu - \delta x - \lambda)^2} \right) = 0 \] (A.26)

\[ r - qx - \frac{\eta_1 h}{(\mu - \delta x - \lambda)^2} - \eta_2 = 0 \] (A.27)

\[ -c + \frac{\eta_1 h}{(\mu - \delta x - \lambda)^2} = 0 \] (A.28)

The complimentary slackness conditions are given by

\[ \eta_1 \left( \beta x - \frac{h}{\mu - \delta x - \lambda} - z \right) = 0 \] (A.29)
\[ \eta_2 (\lambda - \Lambda) = 0 \quad (A.30) \]

Solving, we eliminate the cases of \([\eta_1 = 0, \eta_2 = 0]\) and \([\eta_1 = 0, \eta_2 > 0]\) as infeasible. The interior solution, \([\eta_1 > 0, \eta_2 > 0]\), provides a critical point solution which leads to negative profits; therefore the firm can do strictly better by not entering the market. The only non-trivial solution is for the firm to capture the full market with consumers utility neutral to the outside option, \([\eta_1 > 0, \eta_2 = 0]\). Solving that case provides the result.

\section*{A.1.8 Proof of Theorem 2.4}

\textit{Proof.} Follows from Proposition 2.4 and structure of Theorem 2.1.

\section*{A.1.9 Proof of Proposition 2.5}

\textit{Proof.} We write the Lagrangian as follows

\[ L = r\lambda - c\mu - q\mu x - k + \eta_1 \left( \beta x - \frac{h}{\mu - \delta x - \lambda} \right) - \eta_2 (\lambda - \Lambda) \quad (A.31) \]

The stability conditions are given by

\[ -q\mu + \eta_1 \left( \beta - \frac{h\delta}{(\mu - \delta x - \lambda)^2} \right) = 0 \quad (A.32) \]

\[ r - \frac{\eta_1 h}{(\mu - \delta x - \lambda)^2} - \eta_2 = 0 \quad (A.33) \]

\[ -q x - c + \frac{\eta_1 h}{(\mu - \delta x - \lambda)^2} = 0 \quad (A.34) \]

The complimentary slackness conditions are given by

\[ \eta_1 \left( \beta x - \frac{h}{\mu - \delta x - \lambda} - z \right) = 0 \quad (A.35) \]

\[ \eta_2 (\lambda - \Lambda) = 0 \quad (A.36) \]

Solving, we eliminate the cases of \([\eta_1 = 0, \eta_2 = 0]\) and \([\eta_1 = 0, \eta_2 > 0]\) as infeasible. The interior solution, \([\eta_1 > 0, \eta_2 > 0]\), provides a critical point solution which leads to negative profits; therefore the firm can do strictly better by not entering the market. The only non-trivial solution is for the firm to capture the full market with consumers utility neutral to the outside option, \([\eta_1 > 0, \eta_2 = 0]\). Solving that case provides the result.
A.1.10 Proof of Theorem 2.5

Proof. Follows from Proposition 2.5 and structure of Theorem 2.1. 

A.1.11 Proof of Proposition 2.6

Proof of Proposition 2.6

We write the monopolist’s optimization problem as

\[
\max_{x \geq 0} \pi = r\lambda - qx \quad (A.37)
\]

\[
\beta x - \frac{h}{\mu - \delta x - \lambda} - z = 0 \quad (A.38)
\]

\[
\lambda \leq \Lambda \quad (A.39)
\]

\[
\lambda \geq 0 \quad (A.40)
\]

\[
\lambda < \mu \quad (A.41)
\]

Solving the first constraint, \( \lambda = \mu - \delta x - \frac{h}{\beta x - z} \). Substituting back into the profit function, we write the profit function as follows:

\[
\pi(x) = r\left(\mu - \delta x - \frac{h}{\beta x - z}\right) - qx \quad (A.42)
\]

Using the second-order condition, we can show that the unbounded profit function is concave in \( x \) because \( \beta x - z > 0 \) for any consumers to join with no expected waiting time.

\[
\frac{\partial^2 \pi(x)}{\partial x^2} = -\frac{2rh\beta^2}{(\beta x - z)^3}
\]

Taking the first-order condition, we can solve for the interior solution as follows

\[
\frac{\partial \pi(x)}{\partial x} = r\left(\frac{h\beta}{(\beta x - z)^2} - \delta\right) - q
\]

Again, noting the condition \( \beta x - z > 0 \), we obtain the single positive root:

\[
x = \frac{z}{\beta} + \sqrt{\frac{rh}{\beta (q + r\delta)}}
\]

The equilibrium arrival rate follows from substitution.

\[
\lambda = \mu - \frac{\delta z}{\beta} - (2r\delta + q)\sqrt{\frac{h}{r\beta (q + r\delta)}}
\]
For the full-market boundary solution, $\lambda = \Lambda$

$$\Lambda = \mu - \delta x - \frac{h}{\beta x - z}$$

This equation will have two roots.

**A.1.12 Proof of Theorem 2.6**

The proof follows from Proposition 2.6.

**A.2 Extensions**

**A.2.1 G/G/1 Queue**

An alternative method for checking the robustness of the wait-time assumptions is to extend our results to general arrival and service time distributions.

Now the expected wait time at Provider $i = 1, 2$ is approximated by Kingman’s formula, $W_i(\lambda_i, \mu_i) = \frac{1}{\mu_i} + \frac{1}{2\mu_i} \left( \frac{\lambda_i}{\mu_i - \lambda_i} \right) \left( CV_{a,i}^2 + CV_{s,i}^2 \right)$ where $CV_{a,i}$ denotes the coefficient of variation of the arrival time distribution which has variance $\sigma_{a,i}^2$, where $CV_{s,i}$ denotes the coefficient of variation of the service time distribution which has variance $\sigma_{s,i}^2$. The formula is exact when arrivals are Poisson, and is known to be very accurate when $\rho_i = \frac{\lambda_i}{\mu_i}$ is close to 1. The Nash equilibrium for the symmetric duopoly game solves the following optimization problem.

$$\max_{x_i \geq 0, \mu_i \geq 0} \pi = r\lambda_i - qx_i - c\mu_i - k_i$$  \hspace{1cm} (A.43)

$$\beta x_i - hW_i(\lambda_i, \mu_i) - z_i = 0$$  \hspace{1cm} (A.44)

$$\lambda_1 + \lambda_2 \leq \Lambda$$  \hspace{1cm} (A.45)

$$\lambda_i \geq 0$$  \hspace{1cm} (A.46)

$$\lambda_i \leq \mu$$  \hspace{1cm} (A.47)

From the above, we can again see that given the non-negativity assumptions, it is impossible for a provider to make a positive profit with homogenous consumers if $r < c$ and $k \geq 0$. The optimal solution for the monopolist is given by the following proposition which generalizes the earlier result for M/M/1.

**Proposition A.1.** If $CV_a^2 + CV_s^2 - 2 \geq 0$, then $\lambda^* = \Lambda$
Appendix B

Cancer Screening

B.1 Proofs of Key Results

B.1.1 Proof of Lemma 3.1

Proof. We will derive the probability of each event by summing up the probabilities of all sequences of the four random variables which correspond to that event. We drop the subscripts on the random variables as all variables will regenerate between screens and follow the same distribution independent of cycle number, time, and patient.

Three of the events are explained rather intuitively. First, we derive the probability of detection via other methods at home.

\[ P(SH) = P(X = \min\{X, L, H\}) \cdot P(Y = \min\{Y, L - X, H - X\} \mid X = \min\{X, L, H\}) \]

Due to the memoryless property

\[ P(Y = \min\{Y, L - X, H - X\} \mid X = \min\{X, L, H\}) = P(Y = \min\{Y, L, H\}) = \frac{\theta}{\tau + \alpha + \theta} \]

Substituting we obtain

\[ P(SH) = \left(\frac{\beta}{\tau + \alpha + \beta}\right) \left(\frac{\theta}{\tau + \alpha + \theta}\right) \]

Second, we derive the probability of a screen for cancer which detects cancer (developed at home).

\[ P(CH) = P(X = \min\{X, L, H\}) \cdot P(H - X = \min\{Y, L - X, H - X\} \mid X = \min\{X, L, H\}) \]

Due to the memoryless property
\[ P ( H - X = \min \{Y, L - X, H - X\} \mid X = \min \{X, L, H\}) = P (H = \min \{Y, L, H\}) \]
\[ \quad = \frac{\tau}{\tau + \alpha + \theta} \]

Substituting we obtain
\[ P (CH) = \left( \frac{\beta}{\tau + \alpha + \beta} \right) \left( \frac{\tau}{\tau + \alpha + \theta} \right) \]

Third, we derive the probability of a screen for cancer which detects no cancer.
\[ P (NC) = P (H = \min \{X, L, H\}) = \frac{\tau}{\tau + \alpha + \beta} \]

Now, we examine the probability of a dying at home which can happen in two ways: 1) the patient dies before getting cancer, 2) the patient gets cancer but dies from other causes before it’s detected (via screening or other method). We assume that since the cancer is relatively asymptomatic, there is no cost for living with the disease undetected.

\[ P (DH) = P (L = \min \{X, L, H\}) \]
\[ + P (X = \min \{X, L, H\}) \cdot P (L - X = \min \{Y, L - X, H - X\} \mid X = \min \{X, L, H\}) \]

Due to the memoryless property
\[ P (L - X = \min \{Y, L - X, H - X\} \mid X = \min \{X, L, H\}) = P (L = \min \{Y, L, H\}) \]
\[ \quad = \frac{\alpha}{\tau + \alpha + \theta} \]

Substituting we obtain
\[ P (DH) = \left( \frac{\alpha}{\tau + \alpha + \beta} \right) + \left( \frac{\beta}{\alpha + \beta + \tau} \right) \left( \frac{\alpha}{\alpha + \tau + \theta} \right) \]
\[ \quad = \left( \frac{\alpha}{\tau + \alpha + \beta} \right) \left( \frac{\tau + \beta + \alpha + \theta}{\tau + \alpha + \theta} \right) \]

\[ \square \]

**B.1.2 Proof of Lemma 3.2**

*Proof.* From Equations (3.2) and (3.3) we have
\[ a_1 = \lambda + a_2 P (NC \mid Screen) \]  
(B.1)

\[ a_2 = a_1 P (Screen) \]  
(B.2)

The probability of having a screen within a cycle is given by

\[ P (Screen) = 1 - P (DH) - P (SH) = \frac{\tau (\tau + \beta + \alpha + \theta)}{(\tau + \alpha + \beta)(\tau + \alpha + \theta)} \]  
(B.3)

The conditional probability of a negative screen given that a screen occurs is given by

\[ P (NC \mid Screen) = \frac{\tau + \alpha + \theta}{\tau + \beta + \alpha + \theta} \]  
(B.4)

Therefore, we rewrite the balance equations as

\[ a_1 = \lambda + a_2 \left( \frac{\tau + \alpha + \theta}{\tau + \beta + \alpha + \theta} \right) \]  
(B.5)

\[ a_2 = a_1 \left( \frac{\tau (\tau + \beta + \alpha + \theta)}{(\tau + \alpha + \beta)(\tau + \alpha + \theta)} \right) \]  
(B.6)

Solving the balance equations we obtain the following expressions for the steady-state input rates of the queueing network:

\[ a_1 = \frac{\lambda (\tau + \alpha + \beta)}{\alpha + \beta} \]

\[ a_2 = \frac{\lambda \tau (\tau + \beta + \alpha + \theta)}{(\alpha + \beta)(\tau + \alpha + \theta)} \]

B.1.3 Proof of Proposition 3.1

Proof. We show the convexity of the objective function in Equation (3.6) straightforwardly using the second order condition.

\[ \frac{\partial^2 TC}{\partial \mu^2} = \frac{2 \beta \tau \lambda b_1}{(\alpha + \beta)(\tau + \alpha + \theta) \left( \mu - \frac{\lambda \tau (\tau + \beta + \alpha + \theta)}{(\alpha + \beta)(\tau + \alpha + \theta)} \right)^3} \geq 0 \]

The denominator must be positive due to fact that the input rate into the screening node is given by \[ a_2 = \frac{\lambda \tau (\tau + \beta + \alpha + \theta)}{(\alpha + \beta)(\tau + \alpha + \theta)} \] and the stability condition \( \mu > a_2 \).
B.1.4 Proof of Theorem 3.1

Proof. We write the first order condition for the optimization problem given in Equation (3.6) as follows.

\[ c - \frac{\lambda \beta \tau b_1}{(\alpha + \beta)(\tau + \alpha + \theta)} \left( \mu - \frac{\lambda \tau (\tau + \alpha + \theta)}{(\alpha + \beta)(\tau + \alpha + \theta)} \right)^2 = 0 \]

Solving for \( \mu \) within the feasible region \( \mu > a_2 \) provides the solution

\[ \mu = \frac{\lambda \tau (\tau + \alpha + \theta)}{(\alpha + \beta)(\tau + \alpha + \theta)} + \sqrt{\frac{b_1 \lambda \tau}{c(\alpha + \beta)(\tau + \alpha + \theta)}} \]

The other results are derived straightforwardly by substituting the optimal capacity into the input rates, the wait time function, and the cost function.

B.1.5 Proof of Proposition 3.2

Proof. We show the convexity of the objective function in Equation (3.16) straightforwardly using the second order condition.

\[ \frac{\partial^2 TC}{\partial h^2} = b_1 \lambda^2 \left( \frac{2 \lambda}{(\mu - \lambda - \frac{\lambda}{b h})^2 \beta^2 h^4} + \frac{2}{(\mu - \lambda - \frac{\lambda}{b h})^2 \beta h^3} \right) \geq 0 \]

Both terms must be positive since \( \mu - \lambda - \frac{\lambda}{b h} = \mu - a_2(h) > 0 \).

B.1.6 Proof of Theorem 3.2

Proof. We write the first order condition for the optimization problem given in Equation (3.16) as follows.

\[ b_1 \lambda \left( 1 - \frac{\lambda}{(\mu - \lambda - \frac{\lambda}{b h})^2 \beta h^2} \right) = 0 \]

Solving for \( h \) within the feasible region \( \mu > a_2(h) \) and \( h \geq 0 \), provides the solution.

\[ h = \frac{\lambda + \sqrt{\lambda \beta}}{(\mu - \lambda) \beta} \]

The other results are derived straightforwardly by substituting the optimal average time at home into the input rates, the wait time function, and the cost function.
B.1.7 Proof of Proposition 3.3

Proof. We show the joint convexity of the cost function in Equation (3.18) using the second order conditions.

We write the Hessian of the cost function as follows.

$$H = \begin{bmatrix} \frac{2\lambda b_1}{(\mu - \lambda \frac{\lambda}{\beta h})^3} & -\frac{2\lambda^2 b_1}{(\mu - \lambda \frac{\lambda}{\beta h})^{3} \beta h^2} \\ -\frac{2\lambda^2 b_1}{(\mu - \lambda \frac{\lambda}{\beta h})^{3} \beta h^2} & b_1 \lambda^2 \left( \frac{2\lambda}{(\mu - \lambda \frac{\lambda}{\beta h})^{3} \beta^2 h^4} + \frac{2}{(\mu - \lambda \frac{\lambda}{\beta h})^{2} \beta h^3} \right) \end{bmatrix}$$

The (joint) convexity with respect to $\mu$ and $\tau$ holds due to the second order conditions as given in the following three inequalities

$$H_{1,1} = \frac{2\lambda b_1}{(\mu - \lambda \frac{\lambda}{\beta h})^3} \geq 0$$

$$H_{2,2} = b_1 \lambda^2 \left( \frac{2\lambda}{(\mu - \lambda \frac{\lambda}{\beta h})^{3} \beta^2 h^4} + \frac{2}{(\mu - \lambda \frac{\lambda}{\beta h})^{2} \beta h^3} \right) \geq 0$$

$$\det (H) = \frac{4b_1^2 \lambda^3}{\beta h^3 \left( \mu - \lambda \frac{\lambda}{\beta h} \right)^5} \geq 0$$

Which all hold due to the stability condition $\mu > a_2 (\tau) = \lambda + \frac{\lambda}{\beta h}$.

B.1.8 Proof of Theorem 3.3

Proof. We write the first order condition for the optimization problem given in Equation (3.18) as follows.

$$\frac{\partial TC}{\partial h} = b_1 \lambda \left( 1 - \frac{\lambda}{(\mu - \lambda \frac{\lambda}{\beta h})^{2} \beta h^2} \right) = 0$$

$$\frac{\partial TC}{\partial \mu} = c - \frac{b_1 \lambda}{(\mu - \lambda \frac{\lambda}{\beta h})^{2}} = 0$$

Solving within the feasible region $\mu > \lambda - \frac{\lambda}{\beta h}$ provides the solution
\[ h = \sqrt{\frac{c}{\beta b_1}} \]

\[ \mu = \lambda + \lambda \sqrt{\frac{b_1}{\beta c}} + \sqrt{\frac{\lambda b_1}{c}} \]

The other results are derived straightforwardly by substituting the optimal capacity and average time at home into the input rates, the wait time function, and the cost function.
Appendix C

Revenue Management

C.1 Proof of Proposition 4.1

Proof. Assume $x_2$ Class 2 patients have already been booked. Assume $D_1 \geq x_1$ and the clinic is considering whether or not to accept the $x_1^{th}$ patient. The marginal expected profit, $\Delta x_1 V_0 (x_2, x_1)$, of accepting an additional patient can be written as:

$$\Delta x_1 V_0 (x_2, x_1) = V_0 (x_2, x_1) - V_0 (x_2, x_1 - 1) = p - h P (Z_2 (x_2) > k - b_1) \quad (C.1)$$

Since the marginal expected profit is decreasing in $x_1$, the clinic accepts Class 1 patients so long as $\Delta x_1 V_0 (x_2, x_1) \geq 0$, i.e. $P (Z_2 (x_2) > k - x_1) \leq \frac{p}{h}$, or until all requests have been booked.

C.2 Proof of Proposition 4.2

Proof. Let $b_1 (x_2) = b_1$ for all values of $x_2$. We can write the expected overbooking cost as $L = h E [Z_2 (x_2) + \min (D_1, b_1 (x_2)) - k]^+$. Note that $h (z + \min (D_1, b_1) - k)^+$ is an increasing, convex function of $z$. From Example 8.B.3 on page 368 of Shaked and Shanthikumar (2007) we know \{Z_2 (x_2), x_2 = 0, 1, 2...\} is stochastically increasing convex. Therefore, $E \phi [Z_2 (x_2)]$ is increasing and convex in $x_2$ for any convex function including $h E [Z_2 (x_2) + \min (D_1, b_1) - k]^+$. Therefore, $V_0 (x_2, b_1 (x_2))$ is concave in $x_2$ since the sum of concave functions is concave. This implies $V_1 (x_2, 0)$ must be concave in $x_2$ since the component-wise maximum of concave functions is concave.

C.3 Proposition 4.3

Proof. By definition, for any two random variables, $R_1$ and $R_2$, and increasing function $\phi$

$$R_1 \preceq_{st} R_2 \iff P (R_1 > t) \leq P (R_2 > t) \forall t \iff E \phi (R_1) \leq E \phi (R_2)$$

Since $h [x - k]^+$ is an increasing, convex function of $x$, we can use stochastic ordering between $Z_2 (x_2) + \min (D_1, s - x_2)$ and $Z_2 (x_2 - 1) + \min (D_1, s - x_2 + 1)$ to determine whether the expected overbooking cost is increasing or decreasing.
\[ Z_2(x) + \min(D_1, s - x_2) \geq Z_2(x_2 - 1) + \min(D_1, s - x_2 + 1) \]
\[ Z_2(x_2) - Z_2(x_2 - 1) \geq \min(D_1, s - x_2 + 1) - \min(D_1, s - x_2) \]  \hfill (C.2)

If \( P(Z_2(x_2) - Z_2(x_2 - 1) > t) \geq P(\min(D_1, s - x_2 + 1) - \min(D_1, s - x_2) > t) \) \( \forall t \) then \( Z_2(x_2) \) is stochastically increasing in \( x_2 \). If the reverse is true, then \( Z_2(x_2) \) is stochastically decreasing in \( x_2 \). (Note: each difference is either 0 or 1.)

For \( t \geq 1 \)
\[ P(Z_2(x_2) - Z_2(x_2 - 1) > t) = 0 = P(\min(D_1, s - x_2 + 1) - \min(D_1, s - x_2) > t) \]

For \( t = 0 \)
\[ P(Z_2(x_2) - Z_2(x_2 - 1) > t) = \alpha_2 \]
\[ P(\min(D_1, s - x_2 + 1) - \min(D_1, s - x_2) > t) = P(D_1 \geq s - x_2 + 1) \]

For \( t < 0 \)
\[ P(Z_2(x_2) - Z_2(x_2 - 1) > t) = 1 = P(\min(D_1, s - x_2 + 1) - \min(D_1, s - x_2) > t) \]

For any increasing function \( \phi \), \( Z_2(x_2) \) is stochastically increasing in \( x_2 \) and \( E\phi(Z_2(x_2)) \) is increasing in \( x_2 \) if \( \alpha_2 \geq P(D_1 \geq s - x_2 + 1) \). \( \square \)

### C.4 Proof of Proposition 4.4

**Proof.** Let superscripts \( A \) and \( B \) denote the two cases, respectively, where the Class 1 booking limit stays constant or decreases by one.

**Case A:** \( b_1(x_2) = b_1(x_2 - 1) \)

\[ \Delta V_A^1(x_2, 0) = p\alpha_2 - h\Delta L^A(x_2, b_1(x_2)) \]

If the \( x_2^{th} \) customer does not attend,

\[ \Delta L^A(x_2, b_1(x_2)) = 0 \]
\[ = E[Z_2(x_2 - 1) + \min(b_1(x_2 - 1), D_1) - k]^+ \]
\[ - E[Z_2(x_2 - 1) + \min(b_1(x_2 - 1), D_1) - k]^+ \]

If the \( x_2^{th} \) customer attends and Class 1 demand is at least \( b_1(x_2 - 1) \),
\[ \triangle L^A(x_2, b_1(x_2)) = E[Z_2(x_2 - 1) + 1 + b_1(x_2 - 1) - k]^+ - E[Z_2(x_2 - 1) + b_1(x_2 - 1) - k]^+ \]
\[ = \sum_{z=k-y(x_2-1)}^{\infty} P(Z_2(x_2) > z) \]
\[ - \sum_{z=k-y(x_2-1)}^{\infty} P(Z_2(x_2) > z) \]
\[ = P(Z_2(x_2) > k - b_1(x_2 - 1) - 1) \]

If the \(x_2^{th}\) customer attends and Class 1 demand is equal to \(d < b_1(x_2 - 1)\),

\[ \triangle L^A(x_2, b_1(x_2)) = E[Z_2(x_2 - 1) + 1 + d - k]^+ - E[Z_2(x_2 - 1) + d - k]^+ \]
\[ = P(Z_2(x_2) > k - d - 1) \]

Conditioning upon these sub-cases we write the marginal expected profit as

\[ \triangle V^A_1(x_2, 0) = p\alpha_2 - h\alpha_2 S(x_2, b_1(x_2)) \quad (C.3) \]
\[ - h\alpha_2 P(Z_2(x_2 - 1) > k - b_1(x_2 - 1) - 1) P(D_1 \geq b_1(x_2 - 1)) \]

where

\[ S(x_2, b_1(x_2)) = \sum_{d=k-x_2+1}^{b_1(x_2-1)-1} P(D_1 = d) P(Z_2(x_2 - 1) > k - d - 1) \quad (C.4) \]

**Case B:** \(b_1(x) = b_1(x_2 - 1) - 1\)

\[ \triangle V^B_1(x_2, 0) = p\alpha_2 - p P(D_1 \geq b_1(x_2 - 1)) - h \triangle L^B(x_2, b_1(x_2)) \]

Conditioning on sub-cases as in Case A, we derive the marginal change in expected overbooks as follows and write the marginal expected profit for Case B as

\[ \triangle V^B_1(x_2, 0) = p\alpha_2 - h\alpha_2 S(x_2, b_1(x_2)) \quad (C.5) \]
\[ - P(D_1 \geq b_1(x_2 - 1)) (p - h (1 - \alpha_2) P(Z_2(x_2 - 1) > k - b_1(x_2 - 1))) \]
Bibliography


Ezekiel J. Emanuel and Steven D. Pearson. It costs more, but is it worth more? *NY Times*, January 3 2012.


Gina Kolata. 50 and ready for colonoscopy? doctors say wait is often long - new york times, December 9 2003.


