Cost and Resource Allocation in Healthcare Delivery Systems

by

Maria Fernanda Bravo Plaza

Submitted to the Sloan School of Management
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Management

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2015

Author ..... Signature redacted .....
Sloan School of Management
August 1, 2015

Certified by.... Signature redacted .....
Retsef Levi
J. Spencer Standish (1945) Prof. of Management, Sloan School of Management
Thesis Supervisor

Accepted by ..... Signature redacted .....
Catherine Tucker
Director, Ph.D. Program, Sloan School of Management

Cost and Resource Allocation in Healthcare Delivery Systems

by

Maria Fernanda Bravo Plaza

Submitted to the Sloan School of Management
on August 1, 2015, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Management

Abstract

This thesis studies contemporary challenges arising at the market, system, and organization levels in the healthcare industry, and develops novel frameworks that allow us to better understand cost and resource allocation for strategic decision making in healthcare settings. The U.S. healthcare industry is going through a massive transformation process due to the increasing industry consolidation, and the implementation of the recently enacted healthcare reform. These changes have completely transformed the incentives in the industry, and traditional practices have become outdated, or are, in general, inadequate to address the new challenges. Our frameworks combine real data and statistical analysis with novel optimization-driven approaches (e.g., linear programming, game theory) that capture the first order aspects of the dynamics of the corresponding markets, systems, and organizations. Overall, this work has relied on collaboration with industry partners in order to identify trade-offs, validate models, and pursue practical innovation and implementation of the proposed frameworks.

In the first part, and motivated by real applications from the healthcare industry, we consider a setting, where one firm provides a service to a second firm that is facing stochastic demand for the service. The changes in the reimbursement system have created new opportunities for business-to-business interactions between healthcare systems and providers. Typical contracts in the healthcare industry are based on a transaction fee per unit of service that is negotiated between the two parties. Unlike traditional product-based 2-echelon supply chains, the two firms have opposing risks with respect to the demand volume. We leverage this insight to design a conceptually simple two-price volume based contract, and analyze it within a game theoretic setting. We show that a two-price contract can optimally ensure risk sharing. Moreover, although the resulting problem is non-convex, we are able to characterize the unique equilibrium contract in closed form for a family of utility functions that captures firms' different risk behaviors, and general demand distributions. Moreover, at equilibrium the new contract has two desirable properties: (1) it allows for better risk reduction (measured by CVaR) for the two firms, and (2) it reduces the uncertainty of the payment transaction.

In the second part, we study the strategic cost and resource allocation in large healthcare delivery networks and how these networks can, efficiently, integrate their operations in order to attain network's welfare objectives. Strategic problems, such as resource allocation, capacity placement, and portfolio of services in multi-site networks, require the correct modeling of network costs, network's welfare objectives and trade-offs, and operational constraints. Traditional practices related to cost accounting, specifically, the allocation of overhead and labor cost to individual activities, as a way to account for the consumption of resources, are not suitable for addressing these challenges. These practices often confound resource allocation and network building capacity decisions. In this part, we develop a general methodological optimization-driven framework inspired by network revenue management models, specifically linear programming optimization, that allows us to better understand network costs and provide strategic solutions to the aforementioned problems. We report the application of this framework on a real case study to demonstrate its applicability and important insights derived from it.

Finally, in the third part of this thesis, we study the nature and sources of variability in surgical activities in a large pediatric hospital. We use machine learning techniques to quantitatively show that surgery time variability is high among pediatric cases and, against common belief, this is poorly explained by surgeon effect or other commonly considered characteristics. Our studies suggest that pediatric surgery time has higher inherent variability making pediatric ORs necessarily more costly and harder to schedule than adult ORs. They must therefore be sourced accordingly. These findings are novel and will be useful in the management of busy pediatric operating theaters. For administrators and policymakers, it provides a basis for understanding some of the added costs inherent in caring for children.

Thesis Supervisor: Retsef Levi
Title: J. Spencer Standish (1945) Prof. of Management, Sloan School of Management

Acknowledgments

Foremost, I would like to thank my advisor Retsef Levi for all the support and guidance he provided me along my Ph.D. journey. During all these years, Retsef has been insightful and encouraging not only academically, but also personally. He always believed in me, and inspired me to overcome my fears and challenges along the way. In addition, Retsef has also been a great professional model. I really admire his principles in the classroom and his high ethical standards, which I would like to uphold as I move forward in my academic career. I am so grateful of having him as my advisor.

I would also like to thank Vivek Farias, and Georgia Perakis for serving as part of my thesis committee, and for supporting me in my goal of becoming a professor. I collaborated with Vivek in Chapter 3 of this thesis, and I also was the teaching assistant for his OM class. I really enjoyed these experiences, and learned a lot by observing his performance in the classroom. Georgia has not only been a co-author of the work in Chapter 2 of this thesis, but also a mentor. She has been caring to a personal level beyond her role as the chair of the OM area, and provided me thoughtful advice anytime I needed. In addition, I would like to thank the faculty at Sloan Itai Ashlagi, and Karen Zheng for their advice and support in the completion of this journey.

Another mentor I would like to thank is Dr. Mike McManus (from Boston Children's Hospital). Mike has been an incredible collaborator and the source of invaluable practical knowledge. He is a co-author of the work included in Chapter 4 of this thesis, and it has been a delightful experience to work with him. I am really grateful for his support, and for enriching my understanding of the healthcare delivery in the U.S. - I have learned so much through him!

I have been very lucky to encounter remarkable collaborators - Gonzalo Romero and Marcus Braun (LGO) - during my time at MIT. Gonzalo is an insightful researcher and I really enjoyed working with him in the project described in Chapter 2. Marcus and I worked together in the project described in Chapter 3. He is an excellent professional, always very approachable, and really helpful.

To the Ph.D. program coordinators Hillary Ross and Sarah Massey - thank you for making the Ph.D. years smooth and socially friendly. I really appreciated all those remainders, and certainly enjoyed every single activity you organized for us. I would also like to thank Ariel Brandner for her assistance; she has provided me help far beyond her responsibilities. Thanks to all of you for making things happen in a timely and flawless way.

This journey would not have been the same without the wonderful friends I made along the way. My ORC friends Andre Calmon, Maxime Cohen, David Fagnan, Paul Grigas, Swati Gupta, Vishal Gupta, Nathan Kallus, John Kessler, Will Ma, Velibor Misic, Allison O'Hair, Yaron Shaposhnik, Joline Uichanco, Phebe Vayanos, and Nataly Youssef. Your friendship has made these years the most memorable time of my student life. Special thanks to Angie King, and Kris Johnson for their endless support. Thank you girls for traveling this journey (and all the other journeys

we actually did!) with me, and thank you for being the closest to a family for me here in the U.S. I would also like to give special thanks to Adam Elmachtoub whose friendship has been invaluable to me. I think this thesis would not have been written if it was not for his encouragement and support - Thank you Adam for believing in me, and for making me believe in myself. To my office mates (and office neighbors): Florin Ciocan, Gonzalo Romero, Yiwei Chen, Leon Valdes, Jason Acimovic, Guillaume Saint-Jacques, and Cecilia Zenteno, thank you for your support and innumerable advices along the way. The short (and long) breaks, and the random conversations and laughs made days in the office much more fun. Special thanks to Cecilia for the gym/zumba/pilates breaks! To my Chilean friends Charlie Thraves, Francisco Unda, Leon, Mey, and Gaston, thank you for reminding me of our roots, language, and silly humor. I would also like to thank my wonderful roommates Rashi Gupta and Allison. Thank you girls for keeping up with me, and for the amazing time we shared together!

Friends and relatives from Chile who have been really supportive throughout the years: Ita, Kathy, Panchi, Javiera, Ines, and Vero. Thank you for your love and for keeping me as part of your life despite the distance. In particular, I would like to thank my aunt Piedad, who left our family in the middle of my Ph.D. journey, for her unconditional love and care - you are always in my thoughts.

Last but certainly not least, I would like to thank my family - Patricia Plaza (mom), Eduardo Bravo (dad), Sebastian Bravo (brother), Matias Bravo (brother), and Georgios Mallas (fiancé) - for their unconditional love, and infinite support. I would definitely not be concluding this journey without them. Georgios has been an incredible partner, supportive, caring, and understanding from the very first day. His every day love and support has been essential for the successful completion of my Ph.D., and for been able to pursue my next goal of becoming a professor - thank you for encouraging and helping me in these challenging times, and for making every day a special one. My brothers have been my companions in life and, as the youngest sister; I have always looked up to them. Matias' free spirit and Sebastian's reasoning character made a perfect personality mix! - Thank you guys! My most profound gratitude is for my parents. I have no words to describe how grateful I am for all the effort that they put to educate me and my brothers. They taught me from the very beginning the importance of education, and have supported me in every goal I have endeavored, even when this implied to be far away from them - thank you for your love and inspiration, the person who I am today, and my achievements, are all because of you.

Fernanda Bravo

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The focus of this thesis is the development of novel optimization-driven frameworks to better understand cost and resource allocations for strategic decision-making in healthcare delivery systems and networks. The U.S. healthcare industry is facing a fundamental transformation due to the changes in policy and regulations introduced by the recently enacted healthcare reform. Providers and systems need to rapidly adapt in this new environment. Hence, rethinking traditional practices, which are often inadequate to address the new challenges, is primordial for surviving in the post-reform environment. Specifically, in collaboration with several industry partners, we develop data-statistics and optimization-driven frameworks to study different challenges arising at the market, system (i.e., network), and organization levels within the healthcare industry.

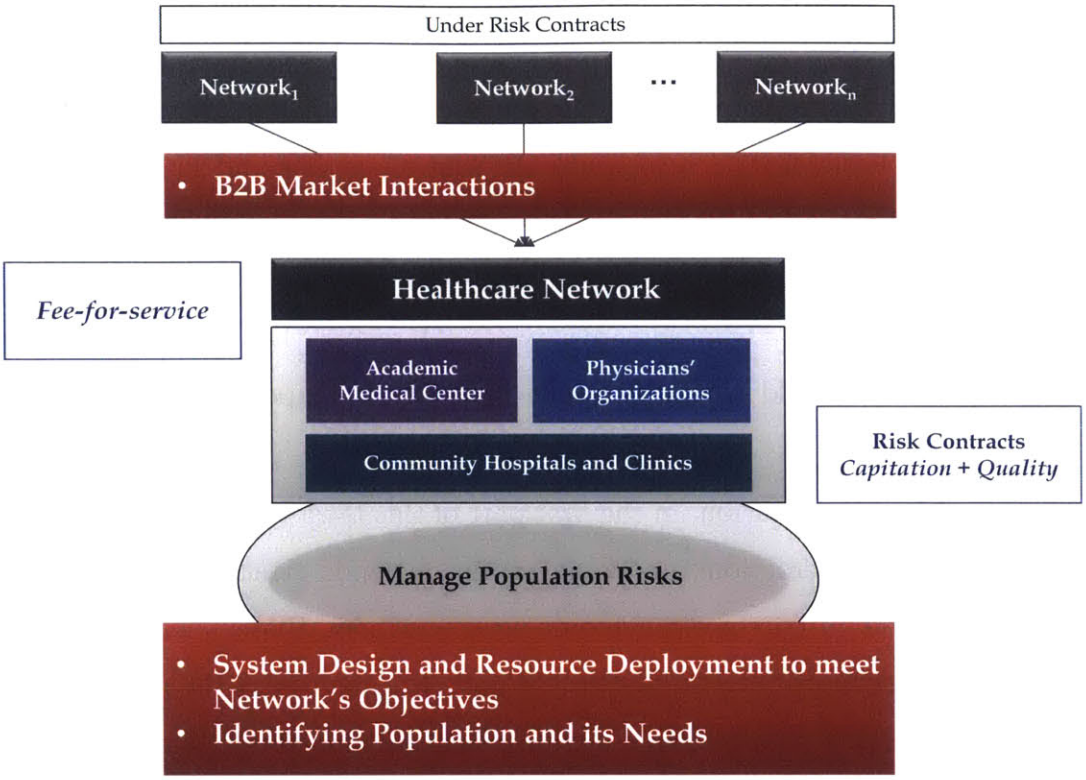
In the last several of decades, the U.S. healthcare industry has undergone a massive trend of consolidation [69]. In the past, academic medical centers, physicians organizations, community hospitals and clinics used to work independently, and operated under a *fee-for-service* type of reimbursement scheme. Under fee-for-service, providers and hospitals are paid (prospectively and independently) based on the volume of services performed. Thus, early consolidation efforts were motivated primarily by the need of building a large physician referral network to gain negotiation power and market share in the industry. In contrast, current consolidation efforts have a much different flavor and motivation. The new healthcare reform was signed into

law in 2010, and with that, universal coverage, cost containment, and quality control initiatives, among many others, assented. One of the major changes introduced by the reform was the migration from the fee-for-service reimbursement system to different risk contracts, which consist of some variation of *capitation* plus quality targets. Under capitation, healthcare systems and providers receive a fixed budget per patient to coordinate and deliver the entirety of the care needs in a specified time period. Thus, healthcare systems and networks are now responsible for managing the risks of their population under a limited budget, moreover, they will be penalized for poor outcomes. Consequently, the need for risk mitigation in the post-reform era has accelerated the industry consolidation trend, and the formation of large healthcare delivery systems and networks.

The contemporary challenges faced by these large systems and networks require the integration of the operations in multi-site networks, the identification of welfare objectives, and the deployment of resources in order to guarantee an appropriate level of access to care in a financially constrained environment. Appropriate levels of access to care are critical to prevent the costly outmigration of patients to other networks. Additionally, and in contrast to the pre-reform environment, these growing systems will have to identify population's needs and manage its risks in a cost effective way. Furthermore, while some of these systems and networks that will manage risk will be able to provide the entire range of required services, a significant number of them will not; these organizations and systems will have to purchase services from other systems. This creates and expands B2B service-based supply chains in the healthcare industry that did not exist before. The most challenging aspect of these business interactions is that the networks that are purchasing services are very sensitive to price since they are under risk contracts themselves. Unfortunately, traditional approaches and practices in the healthcare industry are often outdated, or are in many cases inadequate to address the new challenges arising in the post-reform environment. As a summary, Figure 1-1 illustrates the new challenges that have motivated this thesis.

Our proposed frameworks are based on a quantitative optimization-driven approach that allows us to capture multiple and important trade-offs and systems in-

Figure 1-1: Summary of new challenges in the healthcare industry.



teractions in healthcare settings. In the first part of the thesis, we study the problem arising at the market level that involves the pricing of referrals services between healthcare systems. Specifically, we study how to better share and mitigate risk in a business-to-business (B2B) service based supply chain via appropriate pricing design. In the second part, we address challenges arising at the system level. In particular, we develop an optimization-driven approach to better understand the cost of service, and guide resource allocation and system capacity building decisions in healthcare delivery networks. Finally, in the third part (organization level), we empirically study sources of variability in operational scheduling of surgical activities in a large pediatric academic hospital. In the following sections we provide a summary of the three parts, which conform chapters 2-4 of this thesis.

1.1 Market level: A risk-sharing pricing contract in B2B service-based supply chains

In collaboration with a large tertiary healthcare system, we studied the pricing of referral services between two healthcare systems under demand uncertainty. Our partner has been recently facing a dilemma related to one of its main referring networks (both a provider and an insurance carrier) requesting a discount on the per unit transaction price of advanced care services. Moreover, the referring system were threatening that it would take its volume to another competitor if no discount is offered.

Motivated by this situation, in the first part of this thesis, we study service-based B2B supply chain interactions under uncertain demand. Traditionally, healthcare and other service industries use single unit-price contracts. Under such contracts, the firm requesting service pays a unit-price per unit of demand that the provider serves. Unlike product-based supply chain settings, demand uncertainty induces opposite risks to the firms. Specifically, large demand results in larger revenues for the service provider, but higher cost for the firm requesting service; the single-price contract ignores these opposing risks. We leverage this insight, and propose a simple yet richer pricing contract that consists of a two-price incremental discount contract. We modeled firms' interaction within a game theory framework where firms are modeled as utility-maximizing agents, and their utility is defined according to the principles of Prospect Utility Theory [55] using the single price contract as a reference point.

In more detail, we consider piecewise linear incremental discount contracts (non-negative and non-increasing marginal cost of service) which in its simplest form corresponds to a two-price contract. Indeed, although the resulting problem is non-convex, we can show the optimality of the two-price contract for a family of utility functions, and general demand distributions. Therefore, the service provider, who decides the structure of the contract, does not need to consider more complex contracts in order to guarantee risk sharing. At equilibrium, the two-price contract results in better risk reduction (measured by the CVaR of the payment transaction) for both players.

In addition, we also show that the transaction payment of the new and standard contracts are in Dilation and Lorenz stochastic order [78]. This structural property implies lower variance and lower coefficient of variation of the overall payment transaction. Interestingly, trade-offs captured by this model are general, and common to other service settings, such as post-sale reparation, maintenance, car insurance, and others. This work is included in the paper [6].

1.2 System level: Optimization-driven framework to understand healthcare networks cost and resource allocation

Healthcare delivery systems and networks are facing new challenges due to the increasing industry consolidation trend, and the changes in policy and regulations introduced by the recently enacted healthcare reform. With the implementation of different risk contracts, networks will be responsible for coordinating care and managing the risks of a growing population under a limited budget. Hence, defining population welfare objectives have become a priority. Ensuring appropriate levels of care access, and avoiding the patient outmigration phenomenon, are also crucial to guarantee quality and continuity of care in the financially constrained post-reform environment. Furthermore, the expansion of networks demands the integration of the operations across a multi-site network in order to make an efficient use of the available resources. It also requires identifying network's welfare objectives and deploying resources across the network according to these global objectives, and not based on individual groups' goals, which has been the common practice until recently. Overall, networks need now, more than ever before, to understand the cost of service, and the cost effectiveness of the care that they provide. Unfortunately, common practices around cost accounting, specifically, the allocation of overhead and labor costs to activities as a way to account for the consumption of resources, are often inappropriate for this, and other strategic purposes. In particular, such cost allocations do not capture the

opportunity cost of resources, and they tend to confound decisions related to resource allocation and capacity expansion.

In the second part of the thesis, we develop a general optimization-driven framework, inspired by network revenue management models, to address strategic and operational problems, such as resource allocation, capacity placement, and designing a portfolio of services in a multi-site network. This work was performed in collaboration with the Department of Surgery of a large healthcare network in Boston, Massachusetts. Contrary to the current cost accounting practice in healthcare organizations, our framework distinguishes between two sources of cost: *(i) network capacity cost* (e.g., labor and overhead) and *(ii) service cost* (e.g., supplies, medicines, medical exams). We noticed that the first source of cost is unlikely to vary with changes in the portfolio of services performed, and used this insight in our model. The optimization model can incorporate different networks' *welfare objectives*, such as maximizing profit, throughput, or access, or minimizing costs, or a combination of any of these. In addition, the model can capture many constraints that reflect the operational aspects of the care delivery process, such as surgeon capabilities, resources capacities, block time allocations, and many more. In particular, we model resources that are shared across the network, e.g., surgeon time, which allows us to capture global trade-offs, and the opportunity cost of resources in a network environment.

Case study: we evaluated this framework in our partner hospital network. The managers were concerned about two main issues. First, there was an imbalanced use of surgical capacity across the hospitals. Surgeons at the main campus were perpetually requesting more operating room (OR) time, however, all ORs were fully booked. On the other hand, community hospitals had reported spare OR capacity. The second issue was increasing leaked demand. The integration between our partner and a large payer allowed us to quantify this leaked demand, which corresponds to the volume of patients that received surgical care outside the network. We calibrated the model for about 60 procedure types, including more than 150 resources at 3 hospital locations. Preliminary results suggest that by allowing surgeons to make use of the spare OR capacity in the community, significant improvements can be obtained in

terms of access, as well as in hospital's bottom line (conservative scenarios result in more than 25% increase). In addition, we use the model to determine priorities for procedures in order to maximize hospital's bottom line, and show how this differs with the traditional way in which hospitals' managers prioritize surgical procedures (profit-base, using the traditional cost allocations). This work is part of paper [4].

1.3 Organization level: The nature and sources of variability in pediatric surgical case duration

Surgery is typically a hospital's most costly activity, and the number of surgical suites is limited, so efficient scheduling of operating room (OR) time is an ever-present goal. Unlike adult hospitals, pediatric hospitals have the additional burden of managing extremes of variability that arise from an especially unpredictable patient population. Wide ranges in patient age, size, weight, and developmental level are superimposed upon an even wider range of pathology. In this setting, it is unclear what variability can be removed, and what can be better managed.

In the third part of this thesis, we study sources of operational variability in the operating room. Specifically, we focus on identifying sources of variability in surgical case time. Case time variability confounds surgical scheduling and decreases access to limited operating room resources. Variability arises from many sources and can differ among institutions serving different populations. A rich literature has developed around case time variability in adults, but little in pediatrics. Thus, we study the effect of commonly used patient and procedure factors in driving case time variability in a large, free-standing, academic pediatric hospital. We analyze over 40,000 scheduled surgeries performed over three years. Patient and procedure factors include patient's age and weight, medical status, surgeon identity, and ICU request indicator. We use conditional inference regression trees to analyze these factors and evaluate their predictive power by comparing prediction errors against current practice. We found that pediatrics case time variability, unlike adult cases, is poorly explained by

surgeon effect or other characteristics that are commonly abstracted from electronic records. This largely relates to the ''long-tailed" distribution of pediatric cases and unpredictably long cases. Surgeon-specific scheduling is therefore unnecessary and similar cases may be pooled across surgeons. Future scheduling efforts in pediatrics should focus on prospective identification of patient and procedural specifics that are associated with and predictive of long cases. Until such predictors are identified, daily management of pediatric operating rooms will require compensatory overtime, capacity buffers, schedule flexibility, and cost. This work is published in [5].

Chapter 2

Risk-Sharing Pricing Contract in B2B Service Supply Chains: An Application to Healthcare

2.1 Introduction

In this chapter we study a general 2-echelon B2B service based supply chain and focus on the issue of risk sharing through the appropriate design of pricing contracts. The research was initially motivated by a collaborative work with an industry partner, a tertiary healthcare network, which was under pressure to provide price discounts on referrals to a smaller network on several tertiary care service lines. Specifically, we propose an alternative contract that better captures the risk averseness of the respective parties. The contract is richer, but conceptually simple. Inspired by the pricing of services in practice, we consider incremental discount contracts (non-negative and non-increasing marginal cost of service), and focus on piecewise linear contracts. Furthermore, we consider the traditional (existing) single price contract as a reference point, and leverage Prospect Utility Theory ([55]) to model the risk behavior of the respective parties. We model the interaction between the parties as a full information Stackelberg game in which one party proposes acceptable risk levels and the other

specifies an alternative contract. Using this, we show that risk sharing can be achieved optimally with a two-price contract. The resulting equilibrium two-price contract is characterized via a tractable optimization model that, for interesting special cases, can be solved in closed form. A major insight we derive is that the new contract results in lower variability in the payment transaction between the two parties, which actually coincides with providing lower risk for both of the parties in the supply chain.

Indeed, the recent changes in the healthcare industry in the U.S. have created an increasing number of networks that assume risk and manage specified patient populations for their entire care under a capitation budget. Many of these systems do not have all the in-house expertise to provide the entirety of care required by patients' needs, and must outsource some of the more specialized care (e.g., tertiary care) from another care provider. Contrary to the pre-reform environment, healthcare networks that are in charge of populations have the legal freedom to contract with other providers (outside the network) in order to cover the population's needs. As a result, networks are able to mitigate risk by controlling their referrals patterns (ACO Toolkit, [35]). Following the transactional pricing approach commonly used in services, the pricing of care services would consist of a negotiated single price, that is paid for each unit of service provided. The main advantage of this type of contract is its simplicity. Moreover, these contracts also match the traditional 'DNA' of the healthcare industry in the U.S., that is used to a fee-for-service environment. However, it does not consider any of the risk-related issues that arise because one of the parties in the supply chain is now under a capitated budget, and the fact that the underlying demand for the purchased services is typically stochastic. In particular, the smaller network, which purchases the service, is typically very risk averse with respect to the scenarios that exceed its capitation budget. On the other hand, the larger network, which provides the service, is typically concerned with guaranteeing a minimum level of revenue to cover its large fixed overhead costs. While the initial motivation of this work comes from a healthcare application, the model that we consider captures general service based supply chains. For instance, the auto repair insurance industry, where the insurance company refers the enrollees to specific car

repair dealers (service provider), and pays some portion of the cost of the service depending on the enrollees' plan. Another example is distribution/mailing/delivery services, where a firm outsources the delivery to a third party company, such as UPS, FedEx, or USPS.

2.1.1 Model framework, and assumptions

We consider a firm requesting service or *service requester* (SR, whom we refer to as 'her') that contracts with a *service provider* (SP, whom we refer to as 'him') to purchase a specified service in a defined time window, whereas the demand for the service is a-priori stochastic with known (common knowledge) distribution. The latter assumption is relatively realistic in the healthcare industry, specifically as more states mandate to make claim information publicly available. Motivated by the healthcare application, we assume that any demand must be satisfied by the SR and the SP. Moreover, the SR operates under a (soft) budget to cover the cost of the service over the duration of the contract, and the SP has large capacity, which for all practical purposes is considered a sunk cost. Within the service B2B supply chain, we consider two types of contracts. The first is a negotiated single price contract, in which the two parties agree on a single price per unit of service, independently of the actual volume of service that will be purchased. (This is the current standard contract in the healthcare industry.) The second contract is a piecewise linear incremental discount contract which, in its simplest form, corresponds to a contract with two prices (high and low) and a single breakpoint. Under a two-price contract, the first (higher) price is charged for each service transaction as long as the total volume does not exceed a predefined threshold, and the second (lower) price is charged for each additional transaction once the total volume exceeds the threshold.

We develop a modeling approach to capture the asymmetry in the core risks that the uncertain demand imposes on the two parties in the supply chain. Under the commonly used single price fee-for-service contract, the SR is primarily concerned with *high demand* scenarios that could lead to payment transfers that exceed its capitation budget, while the SP is mostly concerned with *low demand* scenarios, in

which its sunk cost is not covered. It is evident that the current single price contract does not address these risks. (One reason for that is that, traditionally, both parties were under fee-for-service payment schemes so that higher demand was beneficial to both.) On the other hand, an incremental discount contract (e.g., two-price contract) can yield some *overpayment* in the low demand scenarios compared to the single price contract, but at the same time, it induces relatively lower payments (compared to the single price contract) under high demand scenarios (*underpayment*).

Specifically, we consider two approaches to model the firms' risk preferences. On the one hand, we incorporate prospect utility functions to model how firms value the new contract as compared to the the current practice. In addition, we also include specific financial risk constraints to control the risk induced by the new contract. These constraints are specified as *acceptable levels of risk* that the new contract must guarantee. Thus, using Prospect Utility Theory ([55]), we model the utility of the respective parties in the supply chain taking the existing single price contract as a reference point. Specifically, we model the SR as having non-decreasing marginal utility in the induced savings (in large demand scenarios), but an increasing marginal dis-utility in the induced overpayment. The intuition behind this modeling choice is that the SR is willing to use *some* of his budget to cover overpayment. However, when the overpayment becomes too large, the contract rapidly loses its attractiveness due to the opportunity cost of the budget. In addition, we assume that the SP is loss averse, hence, revenue losses under large demand scenarios are perceived more negatively than equivalent gains. By comparing against the commonly used standard contract, we implicitly account for the *status quo bias* that decisions makers often experience in practice, [77]. This bias suggests that decisions makers are inclined to choose current practices over other, equally beneficial, alternatives.

In order to capture the SR acceptable financial risk levels, we use the expected payments and the Conditional Value at Risk (CVaR). This metric has gained significant attention, for example, in the design of insurance policies and portfolio optimization (see [75, 62]), due to its properties as a coherent risk measure. Specifically, the SR is at risk for high cost, and we measured this risk by the CVaR. In our context, the

CVaR has an intuitive interpretation because it coincides with the expected payments above the specified budget. Thus, since the SR must refer all demand, she is particularly concerned with large demand because additional funds would be required. The CVaR captures exactly this risk.

Finally, we model the interaction between the parties as a full information, non-cooperative Stackelberg game. Specifically, the SR will act as the leader and the SP as the follower. This framework has been widely used to model the well-studied seller-buyer interaction in various product-based supply chain settings (e.g., [37, 94]). In the first stage of the game, the SR decides on the acceptable risk levels that the new contract should guarantee. Then, in the second stage, the SP decides the parameters of the new contract (e.g., prices and breakpoints), such that this satisfies the SR's requirements. The players decide their parameters sequentially by maximizing their own individual utility. We believe that the dynamics of this styled game are a reasonable approximation to negotiation dynamics in the healthcare industry.

2.1.2 Contributions

Novel modeling framework to analyze service supply chain contracts. We develop a novel game theoretic based framework that captures the main trade-offs in a service based supply chain in which the parties experience opposing risks due to demand uncertainty. Moreover, we allow the parties to have different risk behaviors and incorporate financial risk budgets explicitly.

Simple pricing contract that allows better risk reduction and less variability in the payments. We leverage the insight that the parties are at risk for opposite extremes of demand into our model to optimally characterize an incremental discount contract. We show that, at equilibrium, a two-price contract can induce optimal risk sharing, hence, there is no need to consider more complex contracts. This new contract reduces the risk of large payments (measured by the CVaR of the single price contract) for the SR, and increases the revenues for the SP in the low demand scenarios, as compared to the single price contract. Furthermore, we show that the

payments induced by the new contract are in *Dilation* and *Lorenz* order ([78]) with the single price contract, which implies the reduction of the overall payment uncertainty.

Closed form solutions for general utility functions, and demand distributions. Although the resulting problem to identify the parameters of the new contract is non-convex, we are able to characterize the unique two-price equilibrium contract in closed form for general utility functions, and demand distributions. This is important because it provides a new contract that can be communicated in practice, and it also facilitates the derivation of additional properties and insights from it.

2.1.3 Literature review

This work relates to different pieces of literature, the product based supply chain and risk management, the service operations management, and the healthcare economics literature.

It is insightful to contrast the setting studied in this paper with the well-studied product based 2-echelon supply chain. Contrary to the product supply chain case, no order quantity or inventory are required in the service setting studied in this paper. Instead, all demand must be served upon request. Furthermore, firms in our setting are affected differently by extremes of demand. In contrast, both parties in the product setting benefit from large demand scenarios, and are risk averse towards low demand realizations. Coordination contracts have focused on how to better share the risk in low demand scenarios, to address the effect of double marginalization, and to align selfish behaviors in order to attain the outcomes of the centralized supply chain. Although extensively used in practice, price-only contracts cannot coordinated capacity investment and inventory decisions in supply chains ([9, 63]). Moreover, the efficiency loss has been quantified in [72]. Alternative contracts have been proposed. For a complete review of the different supply chain settings and alternative contracts see [8], and [47]. For instance, quantity discount contracts have been studied extensively for the coordination of specific supply chain environments, see [31, 99], and [7].

In our setting, on the other hand, the supply chain dynamics have a zero-sum nature whereas the notion of a centralized system is not relevant, hence, we are not interested in comparing against it. Moreover, the parties have opposing risks to high and low demand scenarios. These fundamental differences make prior results on product supply chain not applicable to our setting.

The conflicting firms' incentives in our setting are similar to those described in [20, 21]. In these, the authors study the B2B interaction between a buyer and a supplier of an indirect material (e.g., chemical substances) that is required for the buyer's production process. The supplier wants to maximize the transaction volume while the buyer wants to minimize the consumption, and parties' efforts drive the needed volume. Authors show that shared-saving contracts, which allow parties to benefit from reducing consumption, can achieve higher profit and reduce resource consumption, however, effort levels are not first best optimal. In a more general setting, and considering general cost-of-effort functions, for some cases, the supplier can implement the first-best contract. Similarly, [52] studies the double moral hazard problem arising in the outsourcing of equipment repair and restoration services when the vendor experiences financial distress. It shows that the firm contracting the service can attain the first-best outcomes by implementing performance based tiered and linear (only when the service provider has high tolerance to financial distress) contracts. [91] studies contracts for the outsourcing of maintenance services when the contractor presents risk aversion to certain repair costs. It shows that channel coordination is not always possible in this setting. Contracts for collaborative services is analyzed by [76]. Our work differs from these service settings in several ways. Firstly, we assume that demand (stochastic) is exogenous and must be served upon request; hence, optimal effort levels and moral hazard are not our concern here. Instead, we focus our attention on risk sharing pricing contracts that can allow parties to balance their asymmetric risks due to demand uncertainty. In addition, we confine firms' decisions to a specific contract structure (for the service provider) and acceptable levels of risk (for the service recipient). Finally, we also consider different firms' risk behaviors (e.g., risk-averse, risk-loving) by incorporating Prospect Utility Theory ideas

and using the current practice as a benchmark.

By incorporating different firms risk behaviors, our work connects to the product supply chain with alternative risk behaviors, and to the risk management literature. According to this literature, there are few approaches to incorporate alternative risk preferences into supply chain settings. The utility based approach captures agents' preferences through non-linear utility functions. The main idea of this approach is that different risk behaviors (e.g., risk aversion, loss aversion, risk loving, etc.) is represented using different (shape) utility functions (e.g., CARA/DARA/IARA or prospect utility functions). For example, see [98] for newsvendor models with alternative risk preferences. Alternatively, the mean-variance approach in risk management (in hedging) is used to incorporate the agents' risk sensitivity, see for example [41, 11, 30], and the technical review [16]. Another approach is the value-based approach, which incorporates a metric of downside risk (e.g., Value at Risk (VaR) or Conditional Value at Risk (CVaR)) either in the objective (e.g., [101]) or in the constraints (e.g., [104]). For general newsvendor models with downside risk considerations see [53]. In our work, we integrate two of these approaches. We model firms' risk preferences by specific utility functions (Prospect Utility Theory), and explicitly include financial risk constraints to model maximum levels of risk that a player is willing to accept.

From a healthcare economics perspective, the analysis of contractual arrangements has mostly focused on studying the incentives of different pricing contracts in terms of efficiency and quality of care. The general setting consists of a purchaser (e.g., government or insurer) and a service provider (e.g., hospitals and physicians), and the main question of interest has been around how different pricing mechanism (e.g., fee-for-service, capitation, cost sharing, payment-by-performance, etc.) can help to balance the efficiency and quality of care trade off under different information settings. The principal-agent framework is the dominant tool for analysis in these studies. For a review of payment systems and incentives in healthcare see [70]. For instance, cost sharing contracts can result in higher quality of care ([34]), however, they can also result in lower production efficiency ([14]). [15] empirically studies the

benefit of implementing cost sharing contracts in the case where there is asymmetric cost information. It shows that for DRGs with hight cost variation, cost saving can result in up to 60% lower cost for the purchaser. The study of cost sharing contract usually assumes universal prices for all providers and lump-sum transfers to ensure participation. Under no lump-sum transfers, [68] studies the optimal price adjustment for hospitals assuming that prices differ due to observable cost differences. It characterizes the conditions under which positive/negative price adjustments to the high/low cost provider are optimal. In a different information setting, [51] characterizes optimal cost sharing contracts under physicians altruism private information. The incentives under pay-by-performance contracts have been studied, for example, in [40, 54, 65, 67]. [40] considers a dynamic principal-agent problem for the interaction between a purchaser of medical services and a specialized provider. Inefficiencies arise because the purchaser cannot observe the provider intensity of care. The optimal outcome-adjusted payment system consists of a pre-payment per patient and a ex-post payment adjusted based on short-term outcomes. This payment system generalizes capitation. [54] also uses the principal-agent framework to model the interaction between a purchaser, who cares about minimizing cost while ensuring a quality performance target, and a service provider, who makes decisions on how to allocate capacity among the different streams of patients. Under adverse selection and moral-hazard, they show that commonly used contracts do not coordinate the system and that this can be achieved by a threshold-penalty performance-based contract.

As we mentioned previously, in this work we are not concerned with agency problems, instead, we are interested in how agents can share the opposing risk that they face due to demand (referrals volume) uncertainty. Moreover, we consider the current practice (single price contract) as given, and propose an alternative contract by explicitly using the current practice as a benchmark. One of the implication of this modeling feature is that information regarding the cost of service is not relevant to our model. Thus, in our setting the SP's main concern is designing a pricing contract that is in compliance with the SR's acceptable levels of risk which will ultimately allow the SP to maintain the referral volume.

2.2 Modeling approach

In this section, we describe the elements and main assumptions of our model. Specifically, we introduce an alternative new contract, we define the firms' utility and risks behaviors, and their risk participation requirements.

2.2.1 Pricing contracts

We start by describing the standard single price contract (current practice) and its induced risk due demand uncertainty from both the service requester and the service provider perspectives. Then, we introduce an alternative new contract, and provide some intuition on how this can mitigate the players' risks against uncertain demand volume. Throughout this paper, we model uncertain demand by a continuous, non-negative random variable D with known (common knowledge) probability density function ($f : R_0^+ \rightarrow R^+$), and cumulative distribution function ($F : R_0^+ \rightarrow [0,1]$). We use the short hand notation $\overline{D}_d = \mathrm{E}[D|D > d]$ and $\underline{D}_d = \mathrm{E}[D|D \leq d]$. The overall payment, under a given contract, is denoted by $C(D)$, and note that the exact amount is a-priori uncertain until the exact demand is realized.

Based on the extensively used fee-for-service, the single price contract is considered as the standard contract between the SP and the SR. Under this, the SR pays a fixed price per unit of demand that the SP serves. This price is negotiated beforehand, i.e., before the uncertain demand is revealed, and remains fixed for the entire duration of the agreement, e.g., usually a year in healthcare settings, regardless of the posterior demand realization. In addition, the overall payment is perceived differently by the two parties: it corresponds to cost for the SR and to revenue for the SP. Given this difference, the SR and the SP perceive the risk associated with demand uncertainty in opposite ways. While the SR is at risk for large demand realizations (which results in large cost), the SP would be better off on those scenarios (large revenue), and the contrary occurs in low demand realizations. By construction, single price contracts put all the risk of extreme demand values on one of the parties. However, alternative pricing contracts can induce different levels of risk on the firms. For instance, a lump

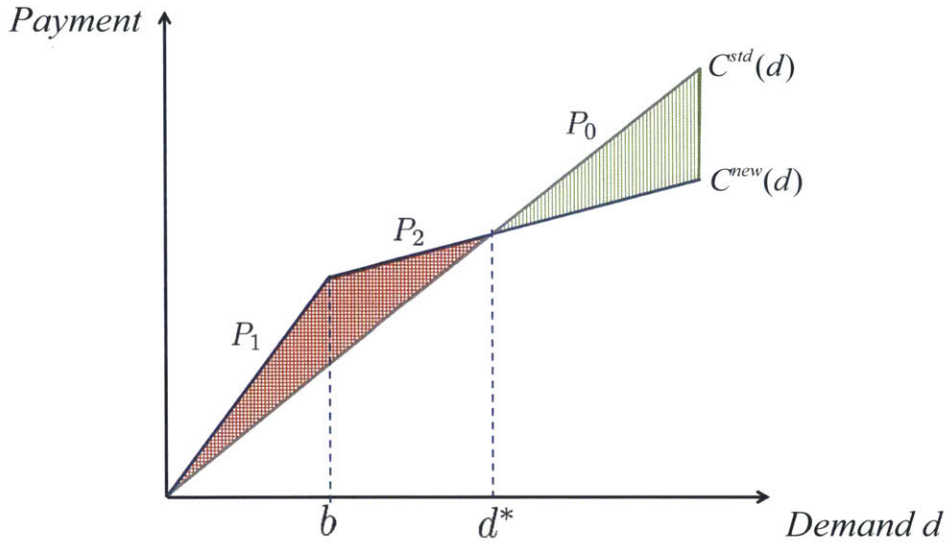
sum money paid upfront, that guarantees service to all demand, puts all the risk of large demand realizations on the SP, who will take care of any demand realization, regardless of its size.

Motivated by the asymmetric risks, we propose an alternative contract that will allow players to share their opposing risks. As commonly observed in the pricing of services in practice, we consider contracts with non-increasing and non-negative marginal cost of service. In particular, we focus on piecewise linear incremental discount contracts because they resemble quantity discount contracts (see [45] for a full description of quantity based discount contracts), and they can be easily communicated and implemented in practice.

For illustration purposes, let us focus on a two-price incremental discount contract, but later on we will show that this is, in fact, enough to guarantee optimal risk sharing. Thus, under a two-price incremental discount contract the first b (breakpoint) units of demand are charged at an initial price P_1, and demand units beyond this threshold are charged at a discounted price $P_2 < P_1$. The design of this new contract, i.e., prices and breakpoint, must be such that, taking into account the parties' different risk behaviors, risk sharing is promoted. In Figure 2-1 we present the payments under the two contracts as a function of the demand. We consider a standard single price contract with fixed price P_0. The overall payment under this contract is $C^{std}(D) = P_0\ D$. Similarly, the total payment under the two-price contract corresponds to $C^{new}(D) = P_1 \min\{b, D\} + P_2(D - b)^+$. By comparing the payments under both contracts, we notice that the new contract induces lower payments in large demand scenarios, i.e., reduce the risk for the SR. However, the new contract also increases payments in low demand scenarios, which are exactly those scenarios in which the SP is at risk for receiving low revenue. Intuitively, the new contract serves as an insurance policy for the SR; it ensures some risk reduction in large demand realizations at expenses of a premium in the expected payment. In order to have this, the prices of the new contract must satisfy $0 \leq P_2 \leq P_0 \leq P_1$. Additionally, we also identify the demand break even point (denoted by d^*), at which the payment of both contracts coincide. Thus, the two-price contract results in *overpayment* in

those demand scenarios to the left of the demand break even point d^*, and results in *underpayment* when demand is above d^*.

Figure 2-1: Comparison of the cost of the standard and two-price contracts.



The prices of the standard (P_0) and new contract (P_1 and P_2), as well as the breakpoint b, and intersection point d^*, are explicitly shown. The area in checked red corresponds to the overpayment, and the area in striped green to the underpayment.

Another advantage of the incremental discount contract is that it provides incentives to the SR to promote more volume to the SP. In the healthcare industry, selective contracting between providers was considered an illegal practice in the past (providers were not responsible for the cost of referrals and these should only be based on patients needs). However, there was still empirical evidence supporting such practices between healthcare providers, see [36, 38, 102], and [60]. In the new regulatory environment, on the other hand, healthcare systems and providers are responsible for cost and have the autonomy to engage in preferential contracts with other systems. Thus, a SR can engage in various contracts with different tertiary care service providers, and can control referrals depending on these contracts. Different contracts

provide different economic incentives, and the proposed incremental discount contract provides incentives for the SR to refer volume for a specific condition to a single SP. This is of tremendous value to the SP who, as previously described, is highly sensitive to low demand volumes.

2.2.2 Valuation of an alternative new contract

As previously described, the new pricing contract induces overpayment and underpayment relative to the standard contract, and the SR and SP value these quantities differently. While the SR perceives overpayment and underpayment as losses and gains, respectively, the SP perceives them as gains and losses, respectively. Moreover, the exact valuation of gains and losses depends on the players' risk behavior. In this paper, we model players' preferences (of a contract) in terms of the changes in the overall payments relative to the status quo, i.e., the single price contract. Thus, we consider utility functions that, based on the players' specific risk behavior, attach a value to the overpayment and underpayment. Utility based on a reference point has been introduced in Prospect Utility Theory [92]. Motivated by empirical anomalies in decision making, and the failure of traditional Expected Utility Theory ([95]) to explain those behaviors, Prospect Utility Theory proposes that agents value alternatives based on gains and losses, relative to the status quo, as opposed to based on their net reported value. In addition to the reference point effect, this theory also considers that agents value gains and losses differently. For instance, loss averse agents tend to discount losses higher than gains. Another relevant feature of this theory is that it explains changes in the valuations of gains and losses as they move farther from the reference point. This behavior can be captured by utilities that have diminishing, increasing, or constant marginal returns, i.e., by introducing utility functions that are concave, convex, or linear in the gains and losses. We incorporate this type of utility functions into our model using the single price contract as a reference point.

In order to model the different risk behaviors, we consider the power family of utility functions which has been previously introduced in [92].

Definition 2.1. Consider a variable $x \in \mathbb{R}$ that represents gains ($x > 0$) and losses ($x < 0$), relative to a predefined reference point. Then, the agent's valuation is defined by

$$u(x) = \begin{cases} x^{\lambda} & x > 0 \\ -\nu\,(-x)^{\theta} & x \leq 0 \end{cases} \tag{2.1}$$

Where $\lambda > 0$ measures the risk aversion in gains, $\theta > 0$ captures the risk aversion in losses, and $\nu > 0$ relates to the sensitivity to losses, i.e., how losses compare against equivalent gains.

For instance, a loss averse agent can be modeled by setting the utility parameters to be $\nu > 1$ and $\lambda \leq \theta \leq 1$. Later on, we will specify the value of these parameters in order to model the specific risk behaviors previously described.

Service Requester (SR). We recall that, in absolute terms, the payment for service is a source of cost for the SR, hence, overpayment and underpayment are perceived as losses and gains, respectively. Thus, the expected utility derived from the new contract can be written depending on whether overpayment and underpayment exist, i.e., depending on whether demand is below or above the demand break even point d^*. Namely

$$\begin{aligned} \mathrm{E}[U_{SR}(C^{std}(D) - C^{new}(D))] =& \mathrm{E}\Big[u_{SR}\Big(C^{std}(D) - C^{new}(D)\Big)\Big|D \leq d^*\Big]F(d^*) + \\ & \mathrm{E}\Big[u_{SR}\Big(C^{std}(D) - C^{new}(D)\Big)\Big|D > d^*\Big](1 - F(d^*)) \end{aligned} \tag{2.2}$$

Generally, the utility is assumed to be increasing in the underpayment, but decreasing in the overpayment. We recall that in our motivating setting, and actually in most service settings, the SR operates under a limited budget (exogenous to our model). This makes the SR very sensitive to large demand scenarios because the total payments might surpass the initial budget allocation. If this is the case, the SR will have to obtain additional funds to cover the cost of service. On the other hand when demand is low, the SR will only spend part of her budget to cover the cost of service. Taking all this together, we consider that the SR's risk preference is driven

by increasing loss aversion ([43]). Under this, although overpayment in low demand scenarios is not desired (dis-utility), it is unlikely that a small overpayment will result in significant financial pressure. However, when the overpayment becomes too large, it cannot longer be financially absorbed and results is larger dis-utility. Hence, due to the opportunity cost of the budget, the SR will experience larger dis-utility as a larger portion of the budget is consumed to cover overpayment. In terms of underpayment, we consider that this generates positive utility with non-increasing marginal returns.

This utility choice was motivated by our conversations with managers in the healthcare industry. The intuition behind is that, given a capitation budget and the current contract, the SR values a reduction in the cost of service (underpayment) in large demand scenarios much more than a unit of unspent budget in low demand scenarios (overpayment). However, if the alternative contract increases payments significantly in the low demand scenarios (i.e., overpayment consumes more of the budget), the valuation may be reversed. We believe this modeling choice applies to general service settings with similar budget considerations.

Service Provider (SP). The overall payment collected from serving the SR demand is a source of revenue for the SP. Therefore, contrary to the SR, overpayment is perceived as a gain and underpayment as a loss. The expected utility is written conditioning on the occurrence of overpayment and underpayment, namely

$$\begin{aligned}\mathrm{E}[U_{SP}(C^{new}(D)-C^{std}(D))] =& \mathrm{E}\Big[u_{SP}\Big(C^{new}(D)-C^{std}(D)\Big)\Big|D\leq d^*\Big]F(d^*)+\\ & \mathrm{E}\Big[u_{SP}\Big(C^{new}(D)-C^{std}(D)\Big)\Big|D> d^*\Big](1-F(d^*))\end{aligned} \tag{2.3}$$

The SP utility is increasing in the overpayment and decreasing in the underpayment. In addition, we consider that the SP predominant risk behavior is loss aversion, meaning that, losses are always perceived worse than equivalent gains. This asymmetry in the valuation of gain and losses is inspired by our motivating example, and represents, for example, the extra effort that the SP will have to incur in order to coordinate activities and deliver the service under large demand scenarios. In addition, and ac-

cording to our motivating example, the SP represents a large and diversified tertiary care provider. Thus, the revenue perceived from this stream of demand is one of many sources, and potential revenue losses, although negatively perceived, will unlikely result in financial pressure for the SP. Therefore, we only consider loss aversion, and not increasing loss aversion, as the primal risk behavior of interest.

2.2.3 Financial constraints: acceptable levels of risk

In this section we define the acceptable levels of financial risk that the new contract must satisfy from the SR's perspective. We emphasize that, in this paper, the focus is on settings where the SR perceives demand as a pure cost source. Moreover, although the SR has a limited budget to cover the cost of service, all demand must be referred to the SP upon realization. Thus, if demand happens to be larger than budgeted, the SR must still refer all the demand to the SP, and obtain additional funds to cover the overall cost of service. Therefore, we consider that, under the standard contract, the SR is particularly concerned with high cost scenarios. In order to capture the SR's sensitivity to large cost, we include a *critical cost value*, above which, the corresponding payments will result in significant financial pressure. The critical cost value is exogenous to our model, and it is related to the SR current contract and financial structure, the capitation budget, as well as the specific practices in the industry.

We define the critical cost value in terms of a confidence level $\beta \in [0, 1]$ of the total cost of service. We recall that, for any contract, the total cost of service $C(D)$ is a-priori uncertain due to the uncertain demand. Thus, for example, if $\beta = 90\%$, the SR has enough budget to cover up to 90% of the lowest cost realizations. If the cost is larger than this (top 10% of cost distribution), the SR will have to obtain additional funds to cover the cost of service. In order to measure the risk of a given contract, we use the Conditional Value at Risk (CVaR). This metric captures the expected payments above the critical cost value and it is of particular interest of the SR since, as we mentioned above, she will have to cover the total cost of service, even if this surpasses the initial budget allocation.

Definition 2.2. For a given contract, consider the induced payment $C(D)$ and a confidence level $\beta \in [0,1]$. The critical cost value corresponds to the Value at Risk of the payment $\Delta_\beta = \text{VaR}_\beta(C(D)) = \min\{\Delta \in \mathbb{R}_0^+ | \mathbb{P}(C(D) > \Delta) \leq 1-\beta\}$. Then, the Conditional Value at Risk is defined by $\text{CVaR}_\beta(C(D)) = \text{E}[C(D)|C(D) > \Delta_\beta]$.

Assuming $C(D)$ continuous and non-decreasing in the demand, the CVaR corresponds exactly to $\text{E}[C(D)|D > d_\beta]$, where $d_\beta = C^{-1}(\Delta_\beta)$, and it is such that, $\mathbb{P}(D > d_\beta) = 1-\beta$. Given the current contract, the demand at risk corresponds to the top $(1-\beta)$-percentile of the demand distribution. Moving forward, we use this latter definition to refer to the risk of a pricing contract and consider β as the demand risk sensitivity. In particular, we are interested in cases where $\beta > F(\bar{D})$, that is, the demand at risk is strictly less than the mean demand.

From the SR perspective, a new contract will result in different levels of risk compared to the standard contract. In order to control for this, we introduce two financial risk constraints that will ultimately allow the SR to select limits on the risk that she is willing to face under a new contract. These constraints were extrapolated from our conversations with healthcare managers and executives, and they, intuitively, act as an insurance policy against demand uncertainty. Thus, let us introduce a *discount* parameter $\alpha \in [0,1]$. The *risk reduction* requirement is modeled as a discount on the risk of the standard contract, that is, the risk of the new contract must be at least $\alpha\%$ lower than the risk of the standard contract; namely

$$\text{E}[C^{new}(D)|D > d_\beta] \leq (1-\alpha)\ \text{E}[C^{std}(D)|D > d_\beta] \tag{2.4}$$

In order to compensate for the risk reduction, and according to the insurance policy interpretation of the new contract, an increase in the expected payments is required. The level of *extra-payment* induced by the new contract must be according to the SR willingness to pay. Thus, let us introduce a *premium* parameter $\gamma \geq 0$, so that the expected payment under the new contract is no more than $\gamma\%$ the payment

under the standard contract; namely

$$\mathrm{E}[C^{new}(D)] \leq (1+\gamma)\ \mathrm{E}[C^{std}(D)] \tag{2.5}$$

Both the discount and premium parameters are chosen by the SR and depend on her specific risk situation and her (soft) budget to cover the cost of service. For instance, if the SR is extremely averse to high cost in large demand scenarios, she might be willing to pay more in order to control the risk. We capture the SR's specific risk limits through the parameters of her utility function.

2.3 Characterization of the new contract

In this section we describe the dynamics between the SR and the SP using a game theoretic framework. Then, we characterize the equilibrium new contract and the acceptable levels of risk demanded by the SR.

2.3.1 B2B dynamics

We employ a full information game to model the interaction between the SR and the SP. Specifically, we use a Stackelberg game where the SR acts as the leader. The motivation for this dynamic comes from what we observed in the healthcare industry; our partner, a tertiary care center, was approached by a network of community hospitals requesting a discount in the price of specialized services, claiming that they would otherwise refer their complex-care cases to a different tertiary care center. In most service industries, and now in healthcare, the firm requesting service chooses where to get service from (or where to refer its demand to), and can usually use this to negotiate better pricing agreements.

As we previously mentioned, we restrict our analysis to incremental discount contracts, and consider that, under the standard single price contract, the SR is at risk for the top $1-\beta\%$ percentile of the demand. The demand sensitivity parameter $\beta \in (0,1)$ is exogenous to our model and known to both players. Thus, in the first

stage the SR chooses her desired level of risk reduction (discount parameter), and the maximum level of extra-payment (premium parameter) that is willing to incur for that risk reduction. Specifically, the SR selects the discount and premium parameters that maximize her utility, and communicates these to the SP. In the second stage, the SP uses this information to characterize the parameters of the new contract (i.e., prices and breakpoint). The players' interaction and their respective decisions take place prior to demand realization. Moreover, we assume that both players are utility maximizers, and optimize their own expected utility as defined in equations (2.2) and (2.3), respectively. Later on, we incorporate a minimum reservation utility for the SP as a way to counterbalance the leader advantage of the SR.

2.3.2 Service Provider's problem

We start by solving the lower level problem. Here, the SP designs a new contract such that his expected utility is maximized. The optimal contract must guarantee the risk reduction and maximum extra-payment requirements (equations (2.4) and (2.5), respectively) quoted by the SR. In concrete, let $\alpha \in [0,1]$ and $\gamma \geq 0$ be the discount and premium parameters, respectively, quoted by the SR in the first stage. Then, the SP optimization problem corresponds to

$$\max_{C^{new}(\cdot)} \mathrm{E}[U_{SP}(C^{new}(D) - C^{std}(D))] \qquad (\mathbf{P^{SP}})$$

$$s.t. \quad \mathrm{E}[C^{new}(D)|D > d_\beta] \leq (1-\alpha)\ \mathrm{E}[C^{std}(D)|D > d_\beta] \quad \text{(Risk reduction, eq. (2.4))}$$

$$\mathrm{E}[C^{new}(D)] \leq (1+\gamma)\ \mathrm{E}[C^{std}(D)] \qquad \text{(Extra-payment, eq. (2.5))}$$

Observe that this is a general formulation since we have not specified the functional form of the new contract, except in that we are implicitly restricting to contracts with non-negative and non-increasing marginal cost of service. This assumption implies that the total payment curves of the new and standard contract cross in at most one point. Moreover, we notice that the above optimization problem is non-convex even if we restrict ourselves to incremental discount contracts (with two or more prices),

and assume a piecewise linear utility function. As specified in equation (2.3), the expected utility is the sum of conditional expectations based on the demand break even point, d^*. This point is an auxiliary variable that depends on the parameters of the contract in a, potentially, non-convex way. For example, for the two-price contract $d^* = \left(\frac{P_1 - P_2}{P_0 - P_2}\right) b$.

Nevertheless, for specific utility functions, the optimal solution (contract) of this general problem has a interesting structure that we can take advantage of. Specifically, we assume that the SP is *loss averse and has a piecewise linear utility function.* This specific utility function is modeled based on the general utility function, equation (2.1), by setting $\lambda_{SP} = 1$, $\theta_{SP} = 1$, and $\nu_{SP} > 1$. In the following result we provide the structural properties of a utility maximizing contract.

Lemma 2.1. *Assume that the SP is loss averse and has a piecewise linear utility function. Then, for given discount and premium parameters $\alpha \in [0, 1]$ and $\gamma \geq 0$, a new contract maximizes the SP utility if and only if*

(i) The risk reduction constraint (2.4) and extra-payment constraint (2.5) are tight

(ii) The demand break even point $d^* = d_\beta$

Moreover, the SP maximum expected utility corresponds to

$$EU^*_{SP}(\alpha, \gamma) = \gamma \mathrm{E}[C^{std}(D)] - (\nu_{SP} - 1)\alpha \mathrm{E}[C^{std}(D) | D > d_\beta](1 - \beta) \tag{2.6}$$

The proof of this Lemma (in Appendix A.1) is straightforward and basically shows that any feasible new contract that does not satisfy (i) and (ii) can be improved. From equation (2.6), we observe that the SP maximum expected utility decreases in the discount level α since lower payments will be received in large demand scenarios. On the other hand, the maximum expected utility increases in the premium γ as larger expected payments will be obtained.

Given our practical motivation, in this paper we are interested in characterizing simple contracts that can be easily communicated between the SP and SR. In particular, we focus on incremental discount contracts. We analyze a two-price contract first,

and later on we will argue that, under certain conditions, there is no need to consider more complicated contracts. We also recall that even if we restrict the problem $\mathbf{P^{SP}}$ to a two-price incremental discount contracts, the resulting optimization problem is still non-convex. However, we can use the structure of the utility maximizing contract in Lemma 2.1 to characterize the optimal two-price contract.

Lemma 2.2. *Under the same assumptions as in Lemma 2.1. For any given discount and premium parameters $\alpha \in [0,1]$ and $\gamma \geq 0$, there is a family of candidate two-price incremental discount contracts that satisfies (i) from Lemma 2.1.*

Moreover, the family of candidate contracts is characterized in closed form as a function of the breakpoint and is such that

- *The first price P_1 is continuous and non-increasing in the breakpoint b, and greater than the price of the single price contract (P_0)*

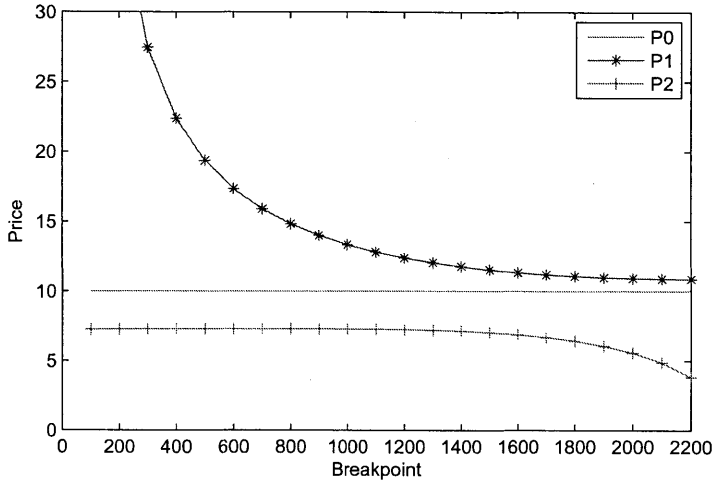
- *The second price P_2 is continuous and non-increasing in the breakpoint b, and smaller than than the price of the single price contract (P_0)*

The proof of Lemma 2.2 is presented in Appendix A.2. The family of candidate contracts $(b, P_1(b,\alpha,\gamma), P_2(b,\alpha,\gamma))$ is parametrized by the breakpoint b (equations (A.3) and (A.4) in Appendix A.2). The set of feasible candidate contracts is reduced to $S(\alpha,\gamma) = \{b | 0 \leq b, 0 \leq P_2(b,\alpha,\gamma)\}$. In the following numerical example, we illustrate how the prices of the family of candidate contracts behave as a function of the breakpoint.

EXAMPLE. Let us consider a log-normal demand distribution with mean 2000 and standard deviation 600. Suppose the SR is at risk for the largest 30% scenarios of the demand distribution ($\beta = 0.7$). The SR would like to reduce the risk by at least 5% ($\alpha = 0.05$). In exchange for the risk reduction, she allows extra-payment of up to 3% ($\gamma = 0.03$). Figure 2-2 shows the prices of he two-price contract as a function of the breakpoint. We observe that as the breakpoint becomes larger, the prices of the family of candidate two-price contracts become smaller. This behavior is intuitive. Specifically, as the breakpoint increases, more units of demand will be charged at the

higher price P_1. Consequently, the two-price contract can support lower prices and still meet the risk reduction and extra-payment requirements.

Figure 2-2: Example of family of candidate two-price contracts.



For any breakpoint, the resulting two-price contract (b, P_1, P_2) satisfies risk reduction and extra-payment constraints with equality. Demand follows a log-normal distribution of mean 2000 and std 600. Discount is 5% and premium 3%. Demand at risk corresponds to the top 30% of the demand distribution.

Interestingly, the two-price contract also captures the well-known two-part tariff pricing contract. This contract consists of an upfront lump-sum fee, as well as a per-unit charge, and it has been extensively studied in product supply chain settings. Specifically, it has been shown that this contract coordinates the interaction (order quantity and profit) between a supplier and a newsvendor retailer, see [10]. Thus, given the closed form expressions of the prices of the two-price contract, we observe that, as the breakpoint approaches zero, the second price reaches its maximum level, which coincides with the per-unit price of the two-part tariff contract that satisfies risk reduction and extra-payment constraints with equality. On the other hand, although the first price of the new contract grows to infinity as the breakpoint approaches zero, the payment of the first b units $(P_1 b)$ converges to a constant. This constant is exactly the lump-sum fee of the two-part tariff contract that meets the risk constraints with equality.

Thus, the equality of the risk reduction and extra-payment constraints helps us to

reduce the search for a two-price utility maximizing contract to the family of candidate contracts. However, even this simplified unidimensional (breakpoint) optimization problem is still non-convex. In order to find the best breakpoint, we use the second property of a utility maximizing contract ((ii) in Lemma 2.1).

Theorem 2.1. *Under the same utility assumptions as in Lemma 2.1 and considering discount and premium parameters* $\alpha \in [0,1]$ *and* $\gamma \geq 0$, *such that*[1]

(a) $\alpha \leq 1 - \frac{d_\beta}{\overline{D}_{d_\beta}}$,

(b) $\gamma \frac{\bar{D}}{\overline{D}_{d_\beta}} \frac{(\overline{D}_{d_\beta} - d_\beta)}{(d_\beta - \bar{D})} \leq \alpha$.

There is a **unique utility maximizing two-price contract**. *The breakpoint of this contract,* $b^* = b^*(\alpha, \gamma)$, *is non-negative and solves*

$$\int_0^{b^*} \left(1 - \frac{t}{b^*}\right) f(t)dt = 1 - \frac{\bar{D}}{d_\beta} - \frac{\gamma}{\alpha} \frac{(\overline{D}_{d_\beta} - d_\beta)\bar{D}}{\overline{D}_{d_\beta} d_\beta} \tag{2.7}$$

Additionally, the prices of this contract are given by

$$\begin{aligned} P_1(b^*, \alpha, \gamma) =& P_0\Big(1 + \frac{\alpha \overline{D}_{d_\beta}}{(\overline{D}_{d_\beta} - d_\beta)}\Big(\frac{d_\beta}{b^*} - 1\Big)\Big) \\ P_2(b^*, \alpha, \gamma) =& P_0 \frac{(1-\alpha)\overline{D}_{d_\beta} - d_\beta}{\overline{D}_{d_\beta} - d_\beta}. \end{aligned} \tag{2.8}$$

Thus, from the SP perspective, a two-price contract attains the maximum possible expected utility, hence, there is no need for considering more complex contracts.

The proof of the Theorem is presented in Appendix A.3. Equation (2.7) determines the breakpoint of the contract such that condition (ii) in Lemma 2.1 is satisfied. Namely, under conditions (a) and (b), the SP can obtain the maximum expected utility by offering a two-price incremental discount contract. These conditions guarantee the feasibility of the two-price contract. In fact, condition (a) is rather general, and

[1]The demand sensitivity parameter β is exogenous and defines the SR's risky scenarios of demand ($D > d_\beta$), where d_β is the β-quantile of the demand. The parameter P_0 is the price of the single price contract. We use the short hand notation $\overline{D}_{d_\beta} = \mathrm{E}[D|D > d_\beta]$ and $\underline{D}_{d_\beta} = \mathrm{E}[D|D \leq d_\beta]$.

corresponds to the maximum risk reduction that the SP can support with any utility maximizing contract that has non-negative and non-increasing marginal cost of service. Condition (b), on the other hand, is specific to the two-price incremental discount contract, and it imposes a minimum risk reduction for any given extra-payment allowance (premium γ) quoted by the SR. Notice that, quoting discount and premium parameters that do not satisfy condition (b) will be detrimental for the SR who will be paying more (than needed) for the desired risk reduction. Therefore, condition (b) is non-restrictive from the SR perspective.

Thus, given the optimality of the two-price contract, we conclude that the SP does not need to consider more complex contracts (with multiple breakpoints) in order to guarantee the SR's acceptable levels of risk. Hereafter, we restrict the analysis to two-price contracts. In the next Corollary, we describe the SP's optimal strategy. We incorporate a reservation utility parameter, $R_{SP} \geq 0$, to capture the SP market power. This modeling feature has been previously used in product supply chain settings to model retailer's opportunity cost or market power (e.g., [63]).

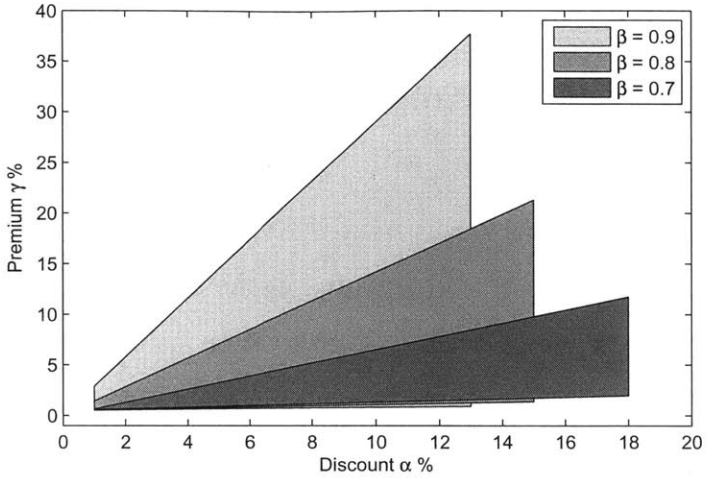
Corollary 2.1. *For any given discount and premium parameters $\alpha \in [0,1]$ and $\gamma \geq 0$, the SP will offer the utility maximizing two-price incremental discount contract, as long as condition (a) and (b) in Theorem 2.1 are satisfied, and the maximum expected utility is at least R_{SP}. The SP's will offer*

$$C^{new*}(\alpha,\gamma) = \begin{cases} \Big(b^*(\alpha,\gamma), P_1(b^*,\alpha,\gamma), P_2(b^*,\alpha,\gamma)\Big) & \text{if } (\alpha,\gamma) \in J \\ & \text{and } EU^*_{SP}(\alpha,\gamma) \geq R_{SP} \\ \text{No contract} & \text{otherwise} \end{cases} \quad (2.9)$$

Where $J = \{(\alpha,\gamma) | 0 \leq \gamma,\ \gamma \frac{\bar{D}(\overline{D}_{d_\beta} - d_\beta)}{\overline{D}_{d_\beta}(d_\beta - \bar{D})} \leq \alpha \leq 1 - \frac{d_\beta}{\overline{D}_{d_\beta}}\}$.

This Corollary is directly derived from Theorem 2.1 (see Appendix A.4). Finally, in Figure 2-3 we provide an illustrative numerical example of the SP's acceptance region, i.e., the discount and premium parameters for which the SP will offer a two-price contract for different levels of demand at risk $(1-\beta\%)$. We observe that as the

Figure 2-3: Example Service Provider's acceptance region.



Demand follows a log-normal distribution with mean 2000 and standard deviation 600. $P_0 = 10$, $\nu_{SP} = 1.2$, $R_{SP} = 100$.

portion of the demand at risk $(1-\beta\%)$ increases (darker region), the feasibility region stretches in the discount direction but shrinks in the premium dimension. Intuitively, when there is more demand at risk the portion of total payments derived from that demand is larger, hence, the allowable discount that the SP can support increases as well. To explain the increase of the premium lower limit, note that as the portion of the demand at risk increases (darker region), less volume will result in overpayment. Thus, the contract requires slightly larger premiums in order to guarantee the SP's reservation utility. At the same time, the premium upper limit is smaller. The reason for this is that when the 'safe' portion of demand is smaller, there is just a limited potential for overpayment.

2.3.3 Service Requester's problem

As the leader, the SR anticipates the new contract that the SP will offer for any discount and premium parameters $\alpha \in [0, 1]$ and $\gamma \geq 0$. As we have identified in section 2.3.2, under the loss aversion and piecewise linear utility assumptions, without loss of generality the SP will offer the utility maximizing two-price contract characterized in equation (2.9). Thus, the SR chooses discount and premium parameters $\alpha \in [0, 1]$

and $\gamma \geq 0$ to maximize her expected utility (given by equation (2.2)), while ensuring the participation of the SP. Namely, we can write the SR's problem as

$$\max_{(\alpha,\gamma)\in J} \mathrm{E}[U_{SR}(C^{new*}(D,\alpha,\gamma) - C^{std}(D))] \qquad (\mathbf{P^{SR}})$$

$$s.t. \quad EU^{*}_{SP}(\alpha,\gamma) \geq R_{SP} \qquad (2.10)$$

Here we consider that the SP participates in the agreement if the maximum expected utility he obtains is, at least, his reservation utility, and the discount and premium parameters satisfy conditions (a) and (b) in Theorem 2.1 (i.e., discount and premium are in J).

As foretold in section 2.2.2, we assume that the SR's risk preferences are driven by an *increasing loss aversion behavior*. Under this assumption, the SR is willing to tolerate small overpayment (losses) in order to reduce her risk in large demand scenarios. However, as the overpayment becomes larger, the disutility from losses rapidly increases. Spending the budget in overpayment is not desirable due to its opportunity cost. The increasing loss aversion behavior has been observed empirically, see [43], where the authors conducted a series of experiments to show that agents experience loss aversion in relation to the size of the losses: agents discount small losses, and tend to emphasize small gains instead. Thus, based on the general utility definition (equation (2.1)), we consider the utility parameters $\theta_{SR} > 1$ and $\nu_{SR} > 0$, small, and for simplicity, we consider $\lambda_{SR} = 1$.

Under the increasing loss aversion behavior, the SR's optimization problem is a-priori hard to solve due to the non-linear dependency of the two-price contract parameters (i.e., breakpoint and prices) on the discount and premium variables. Specifically, we recall that the breakpoint of the contract is the solution to the non-linear equation (2.7) which is a function of the discount and premium variables. Nonetheless, we are able to identify the structure of the optimal solution, and characterize it, in closed form, for special cases. The next result states the conditions under which the problem $\mathbf{P^{SR}}$ admits solution, and it shows that this solution is, in fact, unique.

Theorem 2.2. *Assume that the SP is loss averse and has a piecewise linear utility*

function, and that the SR has an increasing loss aversion behavior. If the parameters of the model satisfy

(a) $\nu_{SP} < 1 + \frac{d_\beta - \bar{D}}{(\overline{D}_{d_\beta} - d_\beta)(1-\beta)}$

(b) $R_{SP} \leq P_0\Big((d_\beta - \bar{D}) - (\nu_{SP} - 1)(\overline{D}_{d_\beta} - d_\beta)(1-\beta)\Big)$

Then, there are unique acceptable risk levels (i.e., discount and premium) that maximize the SR utility.

The proof of Theorem 2.2 is presented in Appendix A.5. Conditions (a) and (b) guarantee that the problem $\mathbf{P^{SR}}$ is feasible. As usual in leader-follower setups, the SP utility constraint is tight at optimality, that is, at equilibrium, the SP obtains exactly his reservation utility. Given this, the premium is expressed in closed form as a function of the discount, see equation (A.12) in Appendix A.5. In the proof of Theorem 2.2, we use this feature to reduce the SR problem to a concave optimization problem where the discount is the single decision variable that is constrained to be in an interval. Thus, the uniqueness of the acceptable risk levels follows directly from the strict concavity of the objective, and we conclude that the game between the SP and the SR has a unique equilibrium in the family of two-price contracts. In particular, when the reservation utility is zero, we can characterize the equilibrium in closed form.

Corollary 2.2. *Under zero reservation utility, the breakpoint of the contract is independent of the discount and premium variables (see equation (A.13) in Appendix A.5). The unique unconstrained maximal discount,* $\tilde{\alpha}$ *corresponds to*

$$\tilde{\alpha} = \frac{\overline{D}_{d_\beta} - d_\beta}{P_0 \overline{D}_{d_\beta}} \left(\frac{(\overline{D}_\beta - d_\beta)(1-\beta)}{\nu_{SR}\theta_{SR}\left(\int_0^{b^*} ((\frac{d_\beta}{b^*} - 1)t)^{\theta_{SR}} f(t)\,dt + \int_{b^*}^{d_\beta} (d_\beta - t)^{\theta_{SR}} f(t)\,dt\right)} \right)^{\frac{1}{\theta_{SR}-1}}$$

Where the breakpoint b^* *is the solution to (A.13) in Appendix A.5.*

For positive reservation utility the problem $\mathbf{P^{SR}}$ is significantly more challenging given that the breakpoint is still a function of the discount variable. Thus, finding

the optimal discount requires solving a system of non-linear equations that, for all practical purposes, can be solved numerically.

2.3.4 Statistical properties of the equilibrium new contract

So far we have established the existence and uniqueness of an equilibrium two-price contract. We now explore the distributional properties of the resulting payments under this equilibrium new contract. For this analysis we consider that the parameters of the model satisfy conditions (a) and (b) from Theorem 2.2.

We first recall that the equilibrium new contract induces larger expected payments than the standard single price contract. Namely,

$$\frac{\mathrm{E}[C^{new}(D) - C^{std}(D)]}{\mathrm{E}[C^{std}(D)]} = \gamma^* = \alpha^*(\nu_{SP} - 1)\frac{\overline{D}_{d_\beta}(1-\beta)}{\bar{D}} + \frac{R_{SP}}{P_0\bar{D}} \tag{2.11}$$

From the SR's perspective, the risk reduction induced by the new contract comes with the price of facing larger overall expected payments. The larger expected payment is required in order to compensate for the SP's degree of loss aversion and market power (reservation utility).

Interestingly, the new contract not only guarantees lower risk for the SR, it also reduces the likelihood of receiving low payments (for the SP) in the lower scenarios of demand. In fact, we can show that, at equilibrium, the new contract reduces the uncertainty in the overall payments. Specifically, we show that the equilibrium payments induced by the new contract and the standard contract satisfy the Dilation and Lorenz stochastic orders. The following variability order definitions are required for this analysis.

Definition 2.3. ([78], definition 3.A.1)

- Convex order: Let X and Y be two random variables such that $\mathrm{E}[\phi(X)] \leq \mathrm{E}[\phi(Y)]$ for all convex functions $\phi : \mathbb{R} \to \mathbb{R}$ provided that the expectations exist. Then X is said to be *smaller than Y in the convex order* (denoted as $X \leq_{cx} Y$).

- Dilation order: $X \leq_{dil} Y$ if $X - \mathrm{E}[X] \leq_{cx} Y - \mathrm{E}[Y]$
- Lorenz order: $X \leq_{Lorenz} Y$ if X and Y are non-negative and $\frac{X}{\mathrm{E}[X]} \leq_{cx} \frac{Y}{\mathrm{E}[Y]}$

These definitions allow us to compare random variables according to their dispersion. Thus, we compare the normalized cost of the new and standard contract, and show that the equilibrium new contract reduces the variability in the overall payments. The next theorem summarizes this result.

Theorem 2.3. *Under the same conditions as in Theorem 2.2, the equilibrium new contract is such that,*

1. $C^{new}(D) \leq_{dil} C^{std}(D)$
2. $C^{new}(D) \leq_{Lorenz} C^{std}(D)$

The proof of this Theorem is in Appendix A.7. These variability orders imply that the variance and coefficient of variation of the payments induced by the new contract are smaller than those induced by the standard contract. The reduction in the payments uncertainty is beneficial for both firms. Indeed, in the proof of Theorem 2.3, we show that the CDF of the payments under the new contract crosses (from below) the CDF of the standard contract in exactly one point. Thus, the equilibrium new contract tends to concentrate the CDF of the payments, so that extreme payments (high and low) are avoided. Therefore, the new contract reduces the likelihood of facing very large and very small payments, which benefits both, the SR and SP, respectively.

2.4 Equilibrium numerical analysis

To study how changes in the model parameters affect the equilibrium contract and acceptable financial risk levels, we analyze the influence of each parameter at the time. In some special cases we can analyze the equilibrium behavior analytically, but in the more general cases the interdependence between the risk levels and the contract

parameters makes the analytical analysis intractable, hence, we provide a numerical equilibrium analysis.

For the numerical analysis, we consider a baseline case which assumes that the demand follows a log-normal distribution with mean 2000 and standard deviation 600. The single price of the standard contract is $P_0 = 10$. The SP's utility function is modeled by a piecewise linear utility function, and the loss aversion sensitivity parameter $\nu_{SP} = 1.2 > 1$. The SP's reservation utility is assumed to be zero ($R_{SP} = 0$). The SR's utility function is assumed to have increasing loss aversion ($\theta_{SP} = 1.6 > 1$), and the sensitivity to losses parameter $\nu_{SR} = 10^{-2}$. The portion of the demand at risk is $1 - \beta = 30\%$.

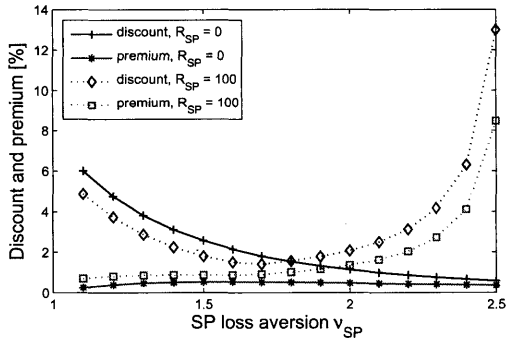
2.4.1 Influence of the SP's loss aversion and reservation utility in the equilibrium contract

The degree of loss aversion (ν_{SP}) and reservation utility (R_{SP}) affect both, the contract parameters (i.e., prices and breakpoint) and the SR's financial risk levels (i.e., premium and discount). Intuitively, higher degree of loss aversion and reservation utility imply that the SP will require larger expected payment in order to offer a new contract that meets the SR desired risk reduction. However, increasing the expected payment is detrimental for the SR. Indeed, our model suggests that if the reservation utility is zero, the SR will avoid the premium increase when dealing with a highly loss averse SP by lowering the level of risk reduction (i.e., by asking for a smaller discount). This strategy allows the SR to maintain an stable premium allowance regardless of the SP level of loss aversion. Equilibrium discount and premium parameters are shown in Figure 2-5(a) plus-sign and star-sign curves, respectively.

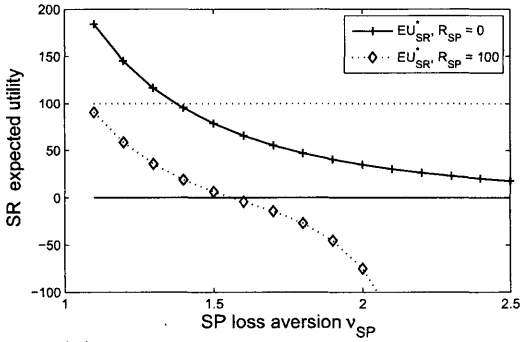
On the other hand, when the reservation utility is non-zero, the equilibrium new contract has to yield enough extra-payment to cover both, the disutility from reducing the risk in high demand scenarios and the higher SP's reservation utility. The model suggests that when the SP degree of loss aversion is low, the level of risk reduction (discount) must decrease and the extra-payment allowance (premium) must increase

relative to the case with zero reservation utility. (plus-sign vs. diamond-sign and star-sign vs. squared-sign curves in Figure 2-5(a) for ν_{SP} small.) Moreover, as the SP degree of loss aversion increases, the SR will have to increase her extra-payment allowance (premium) significantly. Interestingly, the premium increase will also allow the SR to increase the risk reduction that she is requesting from the new contract (diamond-sign curve in Figure 2-5(a)).

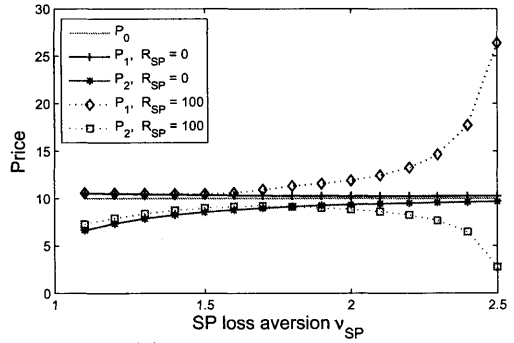
Figure 2-4: Influence of the SP's degree of loss aversion and reservation utility on the equilibrium.



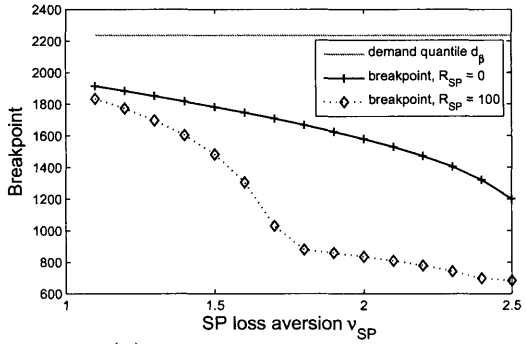
(a) Equilibrium discount and premium parameters.



(b) Utility of the SR at equilibrium.



(c) Equilibrium prices.



(d) Equilibrium breakpoint.

The risky demand scenarios correspond to the top 30%. The SR utility parameters are $\theta_{SR} = 1.6$, $\nu_{SR} = 10^{-2}$.

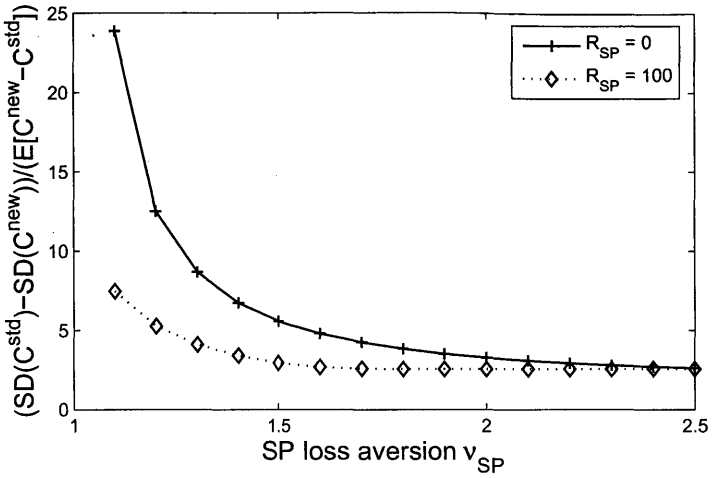
The SR expected utility at equilibrium is shown in Figure 2-5(b). Higher degree of loss aversion and reservation utility reduce the maximum utility that the SR can obtain with the new contract. In particular, when the reservation utility is zero, that is, all the market power is on the SR, the SR can obtain positive utility (plus-sign curve) even if the SP is highly loss averse. However, when the SP has some market

power (diamond-sign curve), and is also highly loss averse, the SR will not be able to obtain positive utility. In such cases, no new agreement will occur between the players and the standard contract will prevail. In practice, this corresponds to the case where the SP dominates the interaction and is not willing to loose revenue in large demand scenarios.

In terms of the prices and breakpoint of the equilibrium two-price contract (Figures 2-5(c) and 2-5(d), respectively), they respond to changes in the utility parameters indirectly through the discount and premium parameters. If the reservation utility is zero, we observe that the first (higher) price of the contract is generally stable, and most of the contract adjustment is done through the second (lower) price and breakpoint. Intuitively, a SP with higher degree of loss aversion and reservation utility requires a contract that induces larger extra-payment (relative to risk reduction). This can be obtained by increasing the second price and lowering the breakpoint of the two-price contract, which is exactly what we observed numerically. Interestingly, the second (lower) price in the case where the reservation utility is non-zero decreases for SPs that are highly loss averse (square-sign curve in Figure 2-5(c)). This behavior is direct consequence of the higher risk reduction (discount) shown in Figure 2-5(a) (diamond-sign curve).

Finally, we also study what the effect of the SP utility parameters is the mean and variability of the new contract total payments. The relative change in the mean of total payments corresponds exactly to the premium parameter (shown in Figure 2-5(a)). The new contract induces higher expected payment in large reservation utility and high loss aversion regimes. In terms of variability of total payments, we notice that the variability reduction offered by the new contract can be *significantly larger* than the increase in the expected payments (ratio is much larger than one). This effect is lessened with the SP degree of loss aversion and reservation utility (see Figure 2-5).

Figure 2-5: Ratio of the standard deviation reduction over expected payment increase.



2.4.2 Influence of the SR's rate of loss aversion and demand at risk in the equilibrium contract

In this section, we analyze the influence of the SR utility parameters in the equilibrium risk levels and contract parameters. Specifically, we study how the equilibrium changes relative to the rate at which the SR becomes loss averse, and to the portion of the demand that is at risk. We interpret the rate at which the SR becomes loss averse as her tolerance to overpayment, thus, a lower rate is equivalent to a high tolerance. We purposely maintain the SR valuation of losses (ν_{SR}) constant, since the influence of this parameter can be directly (and intuitively) analyzed from the closed form solution of the discount parameter detailed in Corollary 2.2.

Let us consider a fixed $\beta = 70\%$, then if the SR tolerance to overpayment in the 'safe' portion of the demand is high (i.e., smaller θ_{SR}), the SR is able to attain the maximum risk reduction (plus-sign curve in Figure 2-7(a)) while paying a very low premium. However, when the SR tolerance to overpayment is low (i.e., larger θ_{SR}), she would prefer to limit the amount of overpayment by reducing her premium (star-sign curve goes to zero). Consequently, the level of discount that the SR can afford with such low premium decreases as well. Similar behavior is observed when the portion of the demand at risk is smaller (i.e., $\beta = 80\%$). However, since in this case a smaller portion of the demand is subject to risk reduction (underpayment), a lower

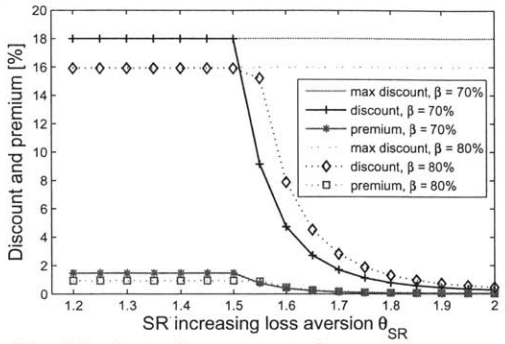
premium is required in order to afford the maximum level of discount (squared-sign vs. star-sign curves). Interestingly, when the tolerance to overpayment is low (i.e., larger θ_{SR}), the SR maintains similar levels of premium, regardless of the portion of the demand at risk, but compensates by asking for a higher discount when there is less demand at risk (compare plus-sign curve vs. diamond-sign curve for larger θ_{SR} in Figure 2-7(a)).

In terms of utility, the maximum expected utility that the SR can obtain with the new contract decreases with the rate at which the SR becomes loss averse (see Figure 2-7(b)). The decay in the utility is due to the discount behavior observed in Figure 2-7(a). Furthermore, we observe that the maximum expected utility is higher when there is more demand at risk ($\beta = 70\%$) and the SR has high tolerance to overpayment. However, the difference in expected utility fades away for SRs with low tolerance to overpayment.

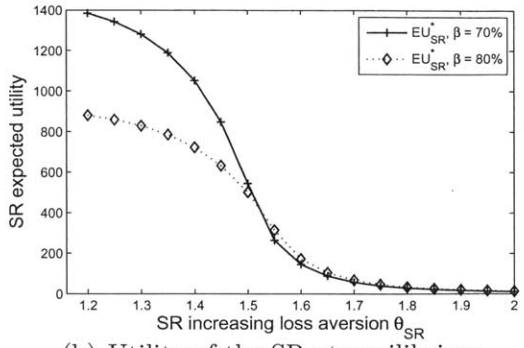
The new contract parameters are adjusted according to the changes in the discount and premium risk levels. Based on the prices behavior shown in Figure 2-7(c), we notice that when the SR has low tolerance to overpayment, she will prefer to continue operating under the single price contract. This behavior can be directly obtained from the prices closed form solution (2.8) in Theorem 2.1. On the other hand, if the SR has high tolerance to overpayment, the equilibrium second price is set to be zero so that the new contract guarantees maximum discount level. In addition, we recall that, under the zero reservation utility assumption, the equilibrium breakpoint is uniquely determined, regardless of the discount and premium values, hence we do not show it here.

Finally, we analyze the effect of the SR tolerance to overpayment and the portion of demand at risk in the expected payment and its variability. As we have already noticed, the equilibrium new contract approaches the single price contract when the SR has a low tolerance to overpayment. The premium parameter corresponds to the relative change in the expected payment induced by the new contract. In Figure 2-7(a), we observe how the premium approaches zero when the overpayment tolerance is low (i.e., larger θ_{SR}). In terms of variability, we observe that the variability reduc-

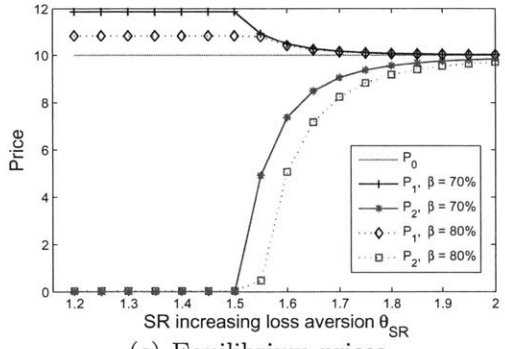
Figure 2-6: Influence of the SR's increasing loss aversion and portion of demand at risk on the equilibrium.



(a) Equilibrium discount and premium parameters.



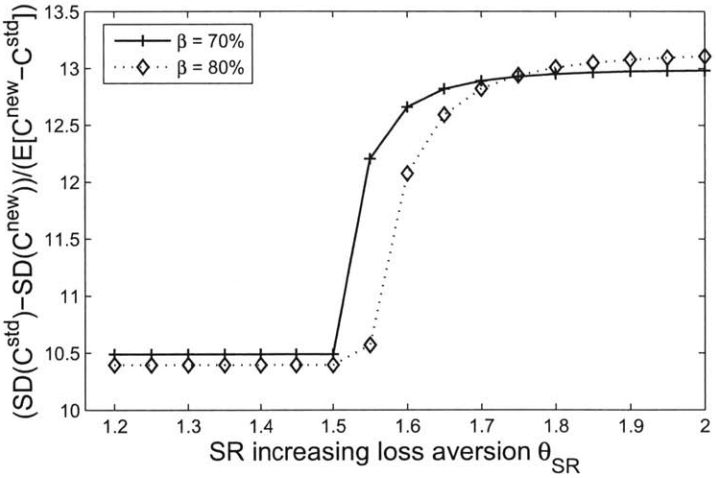
(b) Utility of the SR at equilibrium.



(c) Equilibrium prices.

The SP loss aversion $\nu_{SP} = 1.2$ and the reservation utility is zero.

Figure 2-7: Ratio of the standard deviation reduction over expected payment increase.



The SP loss aversion $\nu_{SP} = 1.2$ and reservation utility is zero.

tion per unit of expected payment increases (see Figure 2-7). The reason for this is that, although the variability reduction offered by the new contract vanishes as the new contract approaches the single price contract, the additional expected payment vanishes at a much faster rate.

2.5 Conclusion

In this work, we studied a 2-echelon service based B2B interaction in which one firm requests a service from another firm, where the demand for that service is stochastic. We showed that better risk sharing between players of opposite risk interests can be achieved by appropriate contracting.

In practice, the pricing of services in most of these applications is based on a per transaction single price contract. Under this contract, firms are at risk for opposite extremes of demand. Specifically, the risk of large demand realization is fully bore by the firm requesting service. On the other hand, low demand realizations result in lower revenues for the firm providing service. In order to balance this risk asymmetry, we consider piecewise linear incremental discount contracts. Specifically, we proposed a simple, yet richer, pricing contract consisting of a two-price contract, and we characterized the equilibrium contract within a game theoretic framework using the single price contract as a benchmark. Moreover, we show that, from the service provider perspective (who decides the structure of the contract), a two-price contract can optimally guarantee risk sharing, thus there is no need in considering more complex contracts.

Although the non-convex nature of the resulting problems, we determined the unique equilibrium contract in closed form for general firms' risk behaviors and demand distributions. At equilibrium, the new contract will result in larger expected payments due to its insurance policy nature; it reduces the risk of large payments for the firm that is paying for the service. However, the new contract ultimately allows better risk reduction for both firms. Intuitively, the equilibrium new contract concentrates the distribution of the payments such that the likelihood of facing extremes

payments (low or high) is decreased.

From a practical perspective, the two-price contract is simple, and can be easily described to firms. Moreover, it has a very intuitive interpretation as an insurance policy against demand uncertainty. Thus, the firm requesting service is able to reduce the risk by paying a specific premium in order to do so, and this premium payment compensates the service provider for offering a lower risk contract.

2.5.1 Future research

One potential extension of this work is the relaxation of the full information assumption. Specifically, the SR has better visibility of the demand than the SP. The question of interest here is how to design a contract that provides the SR with the incentives to reveal the true demand distribution? Alternatively, one can also study how to incorporate preventive measures that the SR may exert in order to partially control the volume of demand. This extension allows us to capture the current preventive care initiatives that many healthcare systems are putting into place in order to avoid tertiary care referrals. Finally, adding competition on the service provider side is also of great interest, specially in the healthcare industry, where tertiary care systems largely compete for the volume of referrals that they can obtain from other small systems and providers.

Chapter 3

Optimization-Driven Framework to Understand Healthcare Networks Cost and Resource Allocation

3.1 Introduction

In this work, we develop a general methodological optimization-driven framework inspired by network revenue management models ([90]), specifically linear programming optimization, that allows us to provide solutions to strategic challenges in healthcare settings. In particular, our framework introduce a new way to understand, and quantify healthcare costs in a network environment. Instead of allocating labor and overhead costs to activities, we directly model resource consumption and capacities, and obtain shadow prices as the opportunity cost of resources from the optimization model. The model can support several decisions related to the allocation of existing resources and portfolio of services across the network, and the identification of growing opportunities (based on leaked demand), and network building. Finally, we report the application of the model to a network of hospitals, and describe the benefits and differences with current practices, and the main insights that the managers and executives derived from it.

Over the last several decades, the U.S. healthcare industry has undergone a massive trend of consolidation, ([69]). This, among other factors, has led to the creation of large healthcare delivery networks, which consist of multiple locations of clinics and hospitals with distinct capabilities (e.g., academic hospitals, physicians organizations, community hospitals, and outpatient clinics). The typical reimbursement system in the industry is *fee-for-service*, where hospitals and providers are paid by the volume of services performed. In this context, early consolidation efforts and motivations were primarily to gain market power and positioning in the industry, and most networks remained decentralized and disintegrated. However, the recent health care reform in the U.S. has fundamentally changed these incentives and motivations. With the implementation of different risk contracts, networks will have to now manage the health of populations of different risks (health needs and financial) profiles. In order to satisfy population's needs and objectives, networks will have to guarantee an appropriate level of access to care, and deliver integrated care in a patient centric manner. In this environment, the *leakage phenomenon*, i.e., outmigration of patients to other healthcare systems and networks, can be problematic for the continuity of care, as well as from a economic perspective. In order to ensure appropriate levels of access to care, networks will have to integrate their operations and deploy resources efficiently across their facilities. Specifically, a typical large system offers various procedures (medical or therapeutic) and services, each requiring a bundle of resources, and incurs costs and collects payments or value with their realization. The strategic design and optimization of these complex healthcare delivery systems is challenging. Strategic problems such as resource allocation, network building and capacity placement, and designing location-specific and overall network portfolio of services, require the correct modeling of network costs, network trade-offs, and operational constraints. Moreover, the related decisions should center on attaining network's objectives, such as profit, access, and throughput, rather than focusing on individual activities and services, as is the common practice for most healthcare organizations. Overall, networks need now, more than ever before, to understand the cost of service, and the cost effectiveness of the care that they provide. Unfortunately, common practices around

cost accounting, specifically, the allocation of overhead and labor costs to activities as a way to account for the consumption of resources, are often inappropriate for this, and other strategic purposes. The problem is that the allocation rules used are somewhat arbitrary, and do not allow for an easy way to understand the true cost of activities and services that are subject to complex interactions. In particular, such cost allocations do not capture the *opportunity cost* of resources, and they tend to confound decisions related to resource allocation and capacity building.

3.1.1 Our framework

We propose a general optimization-driven approach, based on linear programming, that can be tailored to address many strategic decisions related to network design, resource allocation, and capacity placement. Specifically, our approach is built around three major concepts. (a) We distinguish between two sources of cost: *network capacity cost* and *service cost. Network capacity cost* captures all the costs related to building network capabilities including physical infrastructure, as well as manpower (e.g., labor and overhead). *Service cost* corresponds to cost directly related to procedures and activities, and includes all the costs that would not be incurred if the service is not performed (e.g., supplies). We note that the network capacity cost is not likely to change significantly because of operational or tactical decisions to perform or not to perform certain procedures or activities in a specific location. Hence, we only include service cost and capture resource consumption of network capabilities directly in our model. This approach stands in marked contrast to the current practices of cost accounting; by directly accounting for the consumption of resources and capacities, our model implicitly captures network building costs through shadow prices. (b) We guide the network design and optimization decisions to maximize a *welfare objective* that considers the entire network and not just a single activity/service or location at a time. (The goal could be any combination of financial metrics, access metrics, and potentially other network level metrics). (c) We separate the network resources into *fixed* and *flexible. Fixed resources* are set in a predetermined location in the network, whereas *flexible* resources can be allocated to various locations as part of the overall

resource allocation.

On a practical level, our approach can support decisions about how to use a multi-site healthcare network in order to meet demand, which procedures to offer at each location, and the corresponding capacity, are just some examples. In terms of resources, our approach can guide decisions on how to allocate limited resources (e.g., operating room time, surgeon time, and specialist time.) across the various procedures and activities in alignment with network's objectives. Additionally, decisions, such as which growth opportunities to pursue, how much capacity to reserve for them, and what surgeon expertise to bring into the network, can be supported as well. Finally, by incorporating specific operations constraints, our framework can also be used to analyze current operations, identify bottlenecks, and evaluate the effect of changes in capacity, payments, and the portfolio of services offered on the network's objectives.

Case study. In collaborative work with a major academic medical center (AMC), we employed this framework to support a range of decisions and analysis of the allocation of surgical resources across its network, which consists of a teaching hospital and two community hospitals. The current model, in which each hospital manages its surgical capabilities independently, has resulted in two undesirable situations: (i) leaked demand that is mostly due to a lack of access for surgical activities, and (ii) the imbalance use of surgical capacity across the network, specifically, there is spare surgical capacity in the community, while the AMC is fully booked. We used our framework to recommend our partner network how to make use of the surgical spare capacity in the community in order to improve access, recapture leaked demand, while maximizing *revenue net of variable cost* (RNVC). This metric corresponds to the difference between total revenue and service cost. The study involved 57 surgical procedure types (across 3 surgical specialties (85% of volume)), and included hundreds of resources across the network. We used historical data to estimate capacities, utilizations, and financial parameters of the model. Among other analyses, we quantified the impact of recapturing leaked demand using existing network capacity. Our estimates suggest that the network can increase RNVC in up to 12%. Furthermore, the reallocation of current volumes, from the main campus to the community, is responsible for an

additional 1-4%, compared to the case with no reallocation. Notice that the increase in networks' bottom line is much greater since the RNVC metric does not include network building cost (fixed cost). Additional managerial insights are derived from these results. The optimal portfolio of procedure types unveils important business information on which procedures should be emphasized/de-emphasized and which should be reallocated to the community. We compared the optimal changes in the portfolio of procedures to the current priorities, which are based on the traditional cost accounting practices; notable differences demonstrated the power of our approach over current practices to the executives and managers at our partner institution.

3.1.2 Contributions

Our work contributes to the understanding of cost and resource allocation, and to the practice of management in healthcare settings in three different ways.

Modeling framework to support strategic cost and resource allocation in a healthcare network environment. We use linear programming to model the interaction between limited resources and the services that consume those resources in a multi-site network environment. In particular, we are able to optimize network's welfare objectives instead of individual hospitals or departments goals. Additionally, we incorporated different resources, fixed and flexible, the former ones belong to specific location, whereas the later ones can be allocated across different locations. All this, allowed us to support strategic decision making (e.g., case mix, capacity expansion, resource allocation) from a network perspective.

New way of quantifying cost of service in a network environment. Contrary to the traditional approach of (arbitrary) allocating labor and overhead costs to individual services and activities as a way to account for resource utilization, we incorporate the consumption of those capacitated resources directly into our model through constraints. This results in a novel way to quantify the cost of service in a network environment, specifically, by incorporating the opportunity cost of resources through the shadow prices of resources constraints.

Case study in partnership with a real hospital network. In collaboration with a large hospitals network, we used our framework to inform surgical portfolio of procedures, resource allocation, and capacity placement in order to recapture leaked demand while maximizing network's RNVC. We estimated the parameters of the optimization model using real financial, inventory and capacities, and resource consumption data. The estimation methodology can be used to guide the implementation of our framework in different healthcare networks. In particular, we compared the model outputs to the traditional way in which surgical procedures are prioritized, and demonstrated how the model decisions differ from it. Insights from these studies were highly valued by the managers and executives in our partner network.

3.2 Current cost accounting and resource allocation practices

In this section, we describe common practices in capacity and resource allocation, and cost accounting in healthcare settings, and discuss their drawbacks for supporting strategic decision making. Specifically, we illustrate how current practices hinder understanding the true cost of services and the cost of building network's capabilities, which can ultimately lead to the sub-optimal decisions.

(a) **Decentralized operations and short-sighted capacity allocations.** For the most part, hospitals and clinics, that form part of a healthcare network, manage their resources and finances independently. Instead of operating as a unified system, hospitals and clinics provide services, and allocate resources, according to their local objectives, even though, these may be in impairment with network's objectives. The problem of this approach is that networks are not able to use resources efficiently, nor to integrate care, and are also not capable of growing organically due to the difficulty in evaluating, and implementing, network level decisions.

Moreover, within each hospital and clinic, the allocation of capacity to the spe-

cific departments, services, and surgeons is, for the most part, based on historical utilizations. This approach ignores the capacity implications of such allocations on other resources across the network. For example, the allocation of OR time to individual surgeons based on past block time utilization does not account for the effect of the resulting allocation on the intensive care unit and floor beds capacity. These units might become overcrowded, which will result in delays, and potentially compromise the quality of care. As [97] and others have argued, allocating resources based on historical utilization does not maximize efficiency; moreover, it can potentially result in higher costs. Alternatively, mathematical programming based methods have been proposed. A large portion of the literature has centered around improving efficiency for the utilization of resources in single hospital departments. For a general review of the last two decades see [50] and [82]. Typical topics include the hospital bed allocation problem (e.g., [42, 44, 57, 103]), and the OR time allocation and scheduling problems (see general review [13]). For example, [71] uses Data Envelope Analysis to tactically decide on the allocation of additional OR time to surgeons (sub-specialties) based on contribution margin. Their analysis considers a single hospital, and ignores limiting resources outside the OR by arguing that no cancellations have been recorded due to limited capacity. [2, 3] report on the use of linear and integer programming for the allocation of OR time to surgeons in a single hospital. The model chooses the optimal case-mix to be within the OR and floor beds capacities, and to preserve current doctors' income level. Recently, [66] proposes a multi-level optimization model to the planning problem in a single hospital, where the goal is to maximize hospital resource efficiency, and improve patients' service level. Although all these models have proven better than ad-hoc allocations, they only approach capacity and resource allocation (single or couple of resources) in a single facility or location.

(b) **Ad-hoc cost allocation.** Estimating the cost of care is a significant challenge for healthcare organizations. Their financial structure is characterized

by large fixed and indirect costs due to the large investment in infrastructure, manpower, and equipment. Hence, a large portion of the network's cost is related to capacity building. This cost is incurred to support various activities across the network, and can not be directly attributable to specific services or activities. The allocation of the these costs is difficult. The difficulty is that the corresponding network capacity is used in a non-homogeneous and case-by-case basis by hundreds (or even thousands) of different activities and services. In practice, healthcare organizations employ principles from activity-based costing to allocate the cost of network capacity to individual services and activities. Activity-based costing, or ABC, is an accounting method to understand and allocate indirect expenses (e.g., labor and overhead) based on resource consumption, see [17, 18, 19]. In healthcare, [12, 39] describe the benefits of using ABC in this specific setting, and [1, 93] describe the implementation of ABC in hospitals. Complementing this stream of literature, [64] shows the value of distinguishing between the cost of used, and unused capacity in a radiology facility, and pinpoint the most common mistakes in the cost allocation of capital investments. More recently, approaches are tailoring ABC to account for the complex resource interactions and consumption patterns in healthcare, e.g., time-driven ABC by [58, 59]. Although cost accounting principles have provided substantial benefits in informing profitability analysis, we argue that strategic network decisions related to the portfolio of services, resource and capacity allocations, should not be blindly based on such cost allocations. The problem is that the resulting service and procedure *allocated cost* obtained from cost accounting practices do not capture the *opportunity cost* of the resources consumed. Furthermore, allocated cost can make services look arbitrary more or less costly depending on arbitrary allocation rules used.

We present the following illustrative example to show how current practices related to cost accounting and capacity allocation can result in sub-optimal decisions, even in a simplified healthcare setting.

ILLUSTRATIVE EXAMPLE. Let us consider an OR department. The capacity of this OR has been increased in one additional room for next year, and the manager has already hired a nursing team to staff the additional room. The corresponding labor cost adds to $\$300K$ per year. Historical performance suggests that the extra operating room will effectively add 1500 operating hours per year. Thus, using ABC, the labor cost per OR hour is \$200. For simplicity, we consider that there are only two procedures types, I and II, and both have the same reimbursement rate, \$1500 per procedure. The specific surgery duration and cost of surgery supplies, for each procedure type, are detailed in the first two columns of Table 3.1. Additionally, there is floor-bed capacity for 800 days, and both procedure types have the same length of stay, 1 day. The manager is interested in deciding which procedure should be prioritized in order to maximize profit.

Table 3.1: Traditional cost allocation.

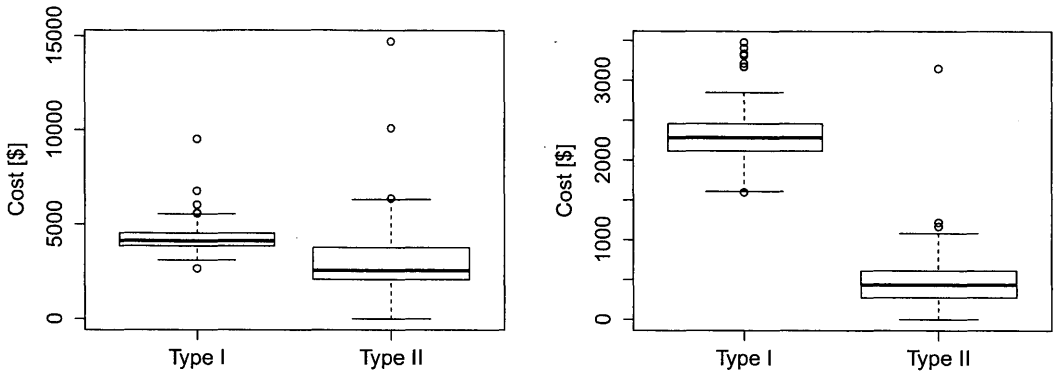
Procedure Type	OR Time [min]	Supplies Cost [\$]	Labor Cost [\$]	Allocated Total Cost [\$]
I	120	100	400	500
II	60	200	200	400

Procedure type I uses twice as much operating time as procedure II, hence the labor cost allocated to the former is twice as large. The cost allocation in Table 3.1 suggests that procedure II is more profitable ($1500 - 400 > 1500 - 500$), hence, the manager would give higher priority to it.

Let us assume that the demand at end of the year is 500 patients type I and 400 patients type II. The manager's priority rule would result in 400 patients for each procedure type accepted. The floor-bed capacity would be fully utilized, and the OR would be 80% utilized. In terms of demand, 100 patients type I would have to be referred somewhere else because the depletion of floor-bed capacity. The total profit would be $\$780K$[1]. Unfortunately, this is not the best the manager could do. At the time of deciding the priorities, the cost of the nursing team (network capacity cost) has been already committed. This cost will be incurred regardless of the type

[1] $400 \times (1500 - 100) + 400 \times (1500 - 200) - 300K$

Figure 3-1: Comparison of 'allocated' cost vs. the cost of the service.



(a) Hospital allocated cost (b) Our service cost.

Note: The cost figures have been scaled for confidentiality. Procedure type I corresponds to Laparoscopic Gastroenterostomy, and procedure type II to Cholecystectomy.

of procedures and volume performed. Thus, priority decisions should be only based on the relevant costs, those that will vary with changes in the case-mix. If instead, we define priorities to maximize *revenue net of variable cost* (revenue - service cost). Then, the optimal priorities are reversed. Without changing the overall volume, the optimal capacity allocation is 500 procedures type I and 300 procedures type II, which results in a profit increase of $\$10K$.

The previous example demonstrates that allocated costs do not capture the interaction between resources and the procedures that consume those resources, and shows how they can mislead capacity allocation decisions. In Figure 3-1, we show a real example of how the allocated cost can distort the relative value of procedures. We compare allocated cost versus service cost (no labor and overhead allocations) for two procedure types that have similar reimbursements rates. When compared based on allocated cost, the mean cost ratio of the two procedures is 1.29, however, when we compare procedures based on service cost, the ratio increases to 4.45, making procedure type II significantly more attractive (cheaper). This difference can significant influence priorities and case-mix decisions.

In reality, capacity and resource allocation decisions are significantly more complex; healthcare networks offer multiple procedures types in several locations. Activities and services consume various resources at different rates, incur different costs while yielding different revenues. Thus, the complexity of this environment makes commonly used ad-hoc approaches inadequate to support strategic network level decisions. Specifically, current practices have a shortsighted view of networks resources and capacities, and completely ignore the interaction among activities and services that compete for these resources, moreover, they do not prioritize network's welfare objectives. This makes the integration of multi-site networks very challenging. In addition, cost accounting practices make almost impossible to understand the actual cost of providing care. The allocation of network capacity costs (e.g., labor and overhead) to individual activities and procedures, as a way to account for resource consumption, confounds the actual cost of providing service. However, network capacity costs are, for the most part, committed regardless of the actual utilization levels, and portfolio of services provided. Therefore, the resulting cost allocations do not represent the actual cost of providing service, and neither capture the opportunity cost of scarce resources. Moreover, when used for strategic decisions, they tend to mix network capacity building and resource and capacity allocation decisions, which most likely results in the inefficient utilization of resources.

3.3 General model

In this section, we introduce our general optimization driven framework, which has been inspired by network revenue management models, [90]. Specifically, we consider a multi-site healthcare network that consists of different hospitals and clinics, several capacitated resources, and various activities and services that can be performed across the different locations. We model the entire network of resources and capacities, and include *network's welfare objectives*, e.g., maximizing profit, throughput, access, or minimizing cost, or any combination of them. In contrast to current cost accounting practices, we only incorporate expenses that are incurred with the real-

ization of the service or activity ([49, 74] use a similar insight for short term case-mix planning), and model the resource consumption of overhead and labor through capacity constraints. As a result, our model dissociates from arbitrary cost allocations and presents a novel way to allocate cost in a network environment, specifically, by looking at the opportunity cost of resources (i.e., shadow prices). In practice, our model can support network building and capacity allocation decisions that take into account the interaction among activities and resources, and it can also facilitate the integration of the operations across the healthcare delivery network.

3.3.1 Elements of the model

In this section, we describe the general elements of our framework.

- **Locations.** These refer to hospitals or clinics located in a specific geographical area. Each location can provide different level of care, for example, academic medical centers can provide a wide range of care, from the most advance and complex type of care to simple and routine visits. Community hospitals and clinics, on the other hand, typically can only treat moderate, simple and routine health conditions.

- **Resources.** These include physical infrastructure (e.g., operating room, floor beds.), equipment, staff, and supplies, whose capacity is limited. We distinguish between *fixed* and *flexible* resources. *Fixed resources* are specific to a particular location and can only be used to deliver care at that location (e.g., surgical supplies, operating room, etc.). *Flexible resources*, on the other hand, are shared across the network, and can potentially be mobilized from one location to another. Thus, these resources can either be allocated among different hospitals (e.g., surgeons who spend two days in the community hospital and three days at the main hospital) or can be used by patients from various locations (e.g., CT scan is at the main hospital but patients from other hospitals can use it too). In addition, we identify a subgroup of resources that are *substitutable*, and they can be safely exchanged for each other (e.g., two general surgeons

with overlapping surgical capabilities).

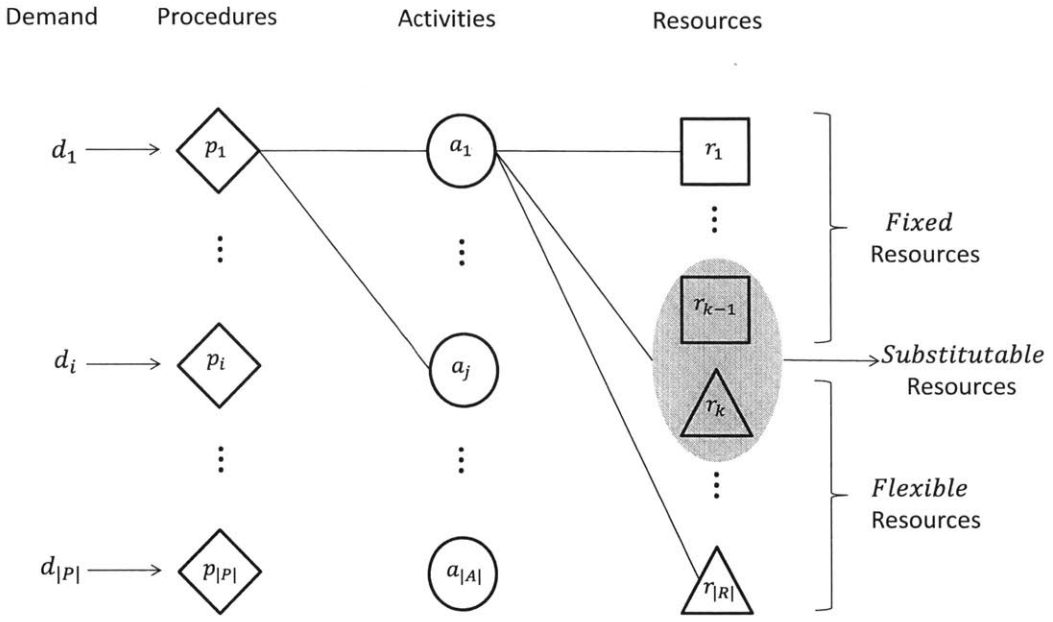
- **Activities.** In order to treat a specific procedure type, a standard set of activities must be executed. For instance, a patient with breast cancer (procedure type) may require a breast surgery (activity 1) and post-radiation or chemotherapy (activity 2). Activities consume a vector of resources, e.g., breast surgery requires a general surgeon for a specific time, an operating room, equipment, and supplies, etc. Moreover, some activities can be demanded by various procedures types, (e.g., chemotherapy) but the resource consumption may be different depending on the specific procedure. Since our model considers volume in an aggregate level, we consider a typical (e.g., mean, median) resource consumption per activity and procedure type. In general, resource consumption is measured in terms of quantity, duration, or time equivalent units (quantity×duration), depending on the specific resource.

- **Procedure types.** We consider a general set of procedures that can be of medical (e.g., diagnosis) or therapeutic (e.g., surgical, rehabilitation) nature.

Locations, resources, activities, and procedure types are connected in the following manner: each procedure type requires the execution of various activities in order to be performed and each activity consumes a certain bundle of resources in order to be executed. Thus, procedures types can be performed at any location as long as the required resources are available for the corresponding activities to be executed. Finally, there is demand for the different procedure types, and each of them generates specific revenue and cost with their realization. Figure 3-2 illustrates the interaction between procedures, activities, and resources.

- **Demand for procedure types.** We consider demand across the network and at each location. Network demand corresponds to the total volume, per procedure type, within the network's geographic area. There is minimum and maximum volume limits that we use to control the demand that the network can serve. These limits aim to capture the extent up to which the network can

Figure 3-2: Diagram of the interaction among procedure types, activities, and resources.



Note: Assuming a specific location, procedure p_1 requires activities a_1 and a_j. Activity a_1 uses fixed resources r_1, flexible resource $r_{|R|}$, and substitutable resource (gray circle) that can be either r_{k-1} or r_k, or a combination of both.

decide on the combination of procedures types to treat. For example, hospitals in the network cannot just focus on the most profitable patients, they must offer a wide variety procedures, even less profitable ones, in order to cover the needs of the population. [61] shows empirical evidence on how merging hospitals tend to redeploy resources to focus on high-profit services but they still maintain a share of non-profitable service lines.

The location-specific demand corresponds to the portion of total volume that must be seen at each specific location. We consider a minimum and maximum demand requirement per location. The goal of this feature is, for example, to control for demand reallocation at a specific hospital or to enforce hospitals to meet minimum volume for specific procedures types. Moreover, this feature

can also be used to ensure diversity in the portfolio of procedures at each location preventing hospitals from converging to uniform services, which has been observed in hospital systems, [32].

As a side note, notice that modeling demand limits requires to quantify existing volumes, as well as the ability to grow demand for specific procedures across the network. The latter task can be significantly hard due to the lack of data. Fortunately, as we will see in the application of this model, the leakage phenomenon will provide us with a good source of information for a realistic estimation of the network's opportunities to grow. Specifically, we can evaluate different scenarios of leaked demand recovery and use that, together with the existing volumes, as a proxy for demand limits.

- **Revenue and cost.** For each procedure type performed, the network collects some revenue, which depends on the reimbursement, and incurs some cost. We consider a typical (e.g., mean, median) revenue and cost per each procedure type. Revenue can be estimated by the payments received by procedure type. Observe that by using typical quantities, we are factoring in differences in payments due to patient insurers, intensity of care delivered, and payment delinquency.

  From the cost side, we distinguish between two sources of cost; the *network building capacity* cost and the *service cost.* The former one includes the costs attributable to infrastructure, equipment, and labor, which determine the operational capacity of the network. At the time one makes resource allocation decisions, these costs are already committed (sunk), and will be incurred regardless of the actual combination of procedures and services performed. The second source of cost corresponds to expenses incurred with each extra unit of service activity or procedure (e.g., supplies, medications, disposable kits, etc.), but otherwise unspent. In terms of the economic cost definition, the first group of costs includes fixed and indirect costs, while the second one only encompasses direct variable costs. Under this distinction, the cost of providing one extra unit

of service only includes the cost of expenses that can be directly attributable to it. This is in contrast to the use of allocated cost resulting from cost accounting principles that most hospitals use in practice; our service cost does not include any cost portions related to network building capacity. Thus, instead of including the cost of network capacity through arbitrary allocations, we directly model the consumption of this capacity and obtain opportunity cost as shadow price as an output of the optimization model.

3.3.2 Mathematical formulation

The mathematical formulation corresponds to a linear programming model which was inspired by the deterministic version of the network revenue management model. In concrete, the model considers a specified time horizon over which it decides on the volume level of each procedure type, at each location, to maximize profitability. Decisions are subject to specific business, and operational constraints, that ensure that a particular combination of procedures can be performed in practice.

Based on the elements previously described, we introduce an index notation, that is, the set of locations, resources, activities, and procedure types, in Table 3.2. In our model, a specific procedure type can only be performed at a given location if all the resources needed for the execution of the corresponding activities are available.

Table 3.2: Definition of sets and indexes.

Set	Notation
Locations	$l \in L$
Procedure types	$p \in P$
Activities	$a \in A$
Fixed resources	$r \in R^{Fix}$
Flexible resources	$r \in R^{Flex}$
Substitutable resources	$r \in R^{Subs}$, $s \in S(r)$

Note: The subset $S(r)$ corresponds to set of substitute resources for resource r. The complete set of resources $R = R^{Fix} \cup R^{Flex}$.

The parameters of the model, that is, the revenue and cost, resource usage, capacities, and demand, are described in Table 3.3.

Table 3.3: Parameters of the model.

Parameter	Notation	Description	Units
Contribution margin	π_{pl}	(Revenue - service cost) obtained by performing procedure type p at location l	Dollars
Resource usage	u_{parl}	Amount of resource r required by activity a to provide procedure type p at location l	Units or time equivalent units
Capacity	c_{rl}, $\tilde{c}_r$, and $\hat{c}_r$	Amount of fixed, flexible, and substitutable resource r available at location l or across the network	Units or time equivalent units
Network demand	Δ_p^-, Δ_p^+	Minimum and maximum network demand for procedure type p	Cases
Individual hospital demand	δ_{pl}^-, δ_{pl}^+	Minimum and maximum demand for procedure type p at location l	Cases

We define two sets of decision variables; x_{pl}: number of cases of procedure type p to be performed at location l, and y_{ral}: amount of substitute resource r allocated to activity a at location l. The second set of variables is necessary in order to assign the correct amount of substitutable resources to activities and to avoid double allocation. Observe that since we are addressing the network optimization from a strategic point of view, the decision variables are assumed continuous, without affecting the interpretation of the optimal solution. The formulation of the network optimization

problem (P) is

$$(P) \underset{\mathbf{x},\mathbf{y}\geq 0}{\text{maximize}} \quad \sum_{l\in L}\sum_{p\in P} \pi_{pl} x_{pl}$$

$$\text{subject to} \quad \sum_{p\in P}\sum_{a\in A} u_{parl} x_{pl} \leq c_{rl} \quad \forall r \in R^{Fix},\, l \in L \tag{3.1}$$

$$\sum_{l\in L}\sum_{p\in P}\sum_{a\in A} u_{parl} x_{pl} \leq \tilde{c}_r \quad \forall r \in R^{Flex} \tag{3.2}$$

$$\sum_{p\in P} u_{parl} x_{pl} \leq \sum_{s\in S(r)} y_{sal} \quad \forall r \in R^{Subs}, a \in A,\, l \in L \tag{3.3}$$

$$\sum_{a\in A}\sum_{l\in L} y_{ral} \leq \hat{c}_r \quad \forall r \in R^{Subs} \tag{3.4}$$

$$\Delta_p^- \leq \sum_{l\in L} x_{pl} \leq \Delta_p^+ \quad \forall p \in P \tag{3.5}$$

$$\delta_{pl}^- \leq x_{pl} \leq \delta_{pl}^+ \quad \forall p \in P,\, l \in L \tag{3.6}$$

In problem (P), constraints (3.1) and (3.2) ensure that the amount of required resources does not exceed the available capacity at each location (fixed resources) and across the network (flexible resources). The left hand side adds up the total amount of resource across procedure types, activities, and locations (only for (3.2)) required to serve the optimal volumes. The right hand side accounts for the total capacity of resources available at each location in (3.1) and across the network in (3.2) in the studied time horizon. The third constraint guarantees the correct allocation of substitute resources across activities and locations. For each substitute resource, the left hand side adds up the total amount of resource required to execute a specific activity across procedure types at a given location. The right hand side corresponds to the total capacity that the model will assign, from the pool of substitutes alternatives, to the specific activity and location. Since substitute resources are modeled as flexible resources, constraint (3.4) guarantees that the allocated capacity across activities and locations does not surpass the network resource capacity for each substitute resource. Finally, constraints (3.5) and (3.6) are demand related. The first one ensures that the network volume for a specific procedure type satisfies a minimum and maximum

demand limit while the second one imposes similar bounds by procedure type and specific location. Observe that the modular structure of the formulation allows us to easily extend this model by incorporating more resources, activities, procedure types, and locations. Moreover, specific constraints to control for operational restrictions, access requirements at each hospital or across the network can also be easily included.

A natural interpretation of the optimal decisions is as capacity budgets; optimal volume levels can be used to derive the corresponding capacity and resource allocation that will be reserved for each procedure type at each location. Moreover, as we will show in the application section, the optimal volumes can be also used to determine utilization of resources, bottlenecks, and to inform acquisition of capacity and the decanting of demand, among many others. In summary, our approach is surprisingly, in stark contrast with current practice; the contrast lies in how one 'prices' the use of resources which must be provisioned well in advance of serving procedures. Existing practice will frequently 'amortize' the real dollar cost of these resources across activities in an ad-hoc fashion based on cost accounting principles. Our approach prices each resource according to the opportunity cost for an additional unit of that resource. This opportunity cost corresponds to the shadow price of the resource and is computed while acknowledging all of the operational constraints one faces in providing services and the demand for those services across the network.

3.3.3 Alternative applications of our framework

A recurrent concern among managers of healthcare systems and networks is *how should they strategically use its limited resources to maximize profitability while upholding the organization's founding mission?* The network optimization model can be used for this and other purposes. For instance, the integration and consolidation of operations, expansion of services, recovery of leaked demand, and improvement of access to care, they all require the optimal management of scarce resources. Our approach addresses these challenges from a strategic and operational (through constraints) perspective. In particular, some of the potential applications of our framework include:

- **Business development applications.** Evaluate how to better use spare capacity across the network; which services should be offered at each location. Determine the priority in which types of leaked procedures should be attacked; which leaked demand is most valuable for the network.

- **Operations Applications.** Determine the extent to which we can meet (or not meet) expected demand with the existing resources, and how capacity should be deployed. Evaluate the financial impact of shifting capacity across the different locations in the network. Identify which resources are limiting and whether it makes financial sense to expand capacity.

In addition to the above applications, the model can also support network building decisions. By incorporating resource consumption and capacities directly, we can derive a comprehensive estimate of the cost of capacity, through shadow prices of the resources. These estimates can be used to evaluate the marginal benefit of capacity investments. Moreover, our model can also inform operational decisions. The capacity and resource allocation in the short term is difficult due to variability in demand arrivals, durations, and resource consumption. Thus, solutions at the aggregate level need to be translated into an operational plan that can guarantee an efficient use of resources. Our model can be easily adapted to account for this operational variability. For instance, in estimating the resource consumption parameters, instead of using the mean or median duration of an activity as the typical consumption, we can simply use higher quantiles. Alternatively, we could also add or modify constraints, for example, by adding buffer capacity to account for the variability in durations, or by reserving capacity for emergency cases, and so on. With these simple adjustments, we will obtain solutions that can better complement and guide the subsequent operational decisions, such as scheduling, staffing, and others.

3.4 Case study

In this section, we describe the application of our framework to a healthcare network consisting of two community hospitals and one Academic Medical Center (AMC). We worked in collaboration with a leadership team from the surgical department at the AMC to estimate the parameters of the model and to study how this network can efficiently use its surgical capabilities to maximize profitability. Specifically, we employ the model to address two fundamental issues; (i) leaked demand, and (ii) the imbalance use of surgical capacity across the hospitals in the network.

Until now, the hospitals in this network have operated independently; each hospital manages its own patient volumes and surgical capabilities. The AMC offers the most advance care, and provides various surgical procedures that are not typically provided in the community setting due to limited resources, surgeonsâĂŹ expertise, and other capabilities. The AMC's surgical volume is highly dependent on referrals derived from affiliated primary care physicians within the network's geographical area. Unfortunately, in the last few years an increasing trend of volume being referred outside the network (leakage phenomenon), and the corresponding revenue loss, has been observed. Lack of timely access to care at the AMC has been conjectured as a driver of leaked demand, specifically, patients being unable to schedule surgical appointments. On the other hand, surgeons at the AMC perpetually request additional operating room time, however, hospital managers claim that operating room capacity is already at maximum levels. At the same time, surgical resources in the community hospitals are not fully utilized. This raised the issue of whether offloading the volume of certain procedures, from the AMC to the community hospitals, would be an effective approach to free up capacity at the AMC, and potentially recover leaked demand, and ultimately create better access across the entire network. Addressing access issues and leaked demand recovery requires a better coordination and visibility of resources across the hospitals in the network. For instance, deciding which surgical procedures to perform at each location depends on several factors. First, not every single procedure type can be provided in a community setting; hospitals handle differ-

ent levels of procedure complexity depending on available resources and capabilities. Furthermore, revenue and cost might differ across locations due to fluctuations in the reimbursement rates based on location and the kind of institution. We used our model to study the feasibility of recapturing leaked demand, and quantify the impact of this on utilizations and network's profitability. To be more precise,

- We use real network's data to estimate the data inputs of the model
- We calibrate the model by imposing 'current' constraints to capture existing state (baseline)
- We perform two analyses: (1) Estimate the value of capacity at the AMC, (2) Quantify the value of recapturing leaked demand using existing network capacity. We compare our insights against a baseline scenario and current practices.

Additionally, we provide insightful information on which procedures to offload to the community, what the best use of spare capacity is, and how this should be allocated to the AMC's surgeons. Finally, we also analyze changes in utilization as a proxy for access improvement.

3.4.1 Data collection and estimation of model parameters

Our work focuses on a family of 57 surgical procedures from three service lines at the AMC (85% of their 2012 volume), and all data entries have been estimated based on 2012-2013 operations. Historical data on surgery duration, length of stay, and required resources were retrieved from the AMC operational and administrative data-bases. Since hospitals IT systems are not integrated, the AMC's data-bases only contain historical records of surgical patients at this location. Thus, data for the community hospitals had to be collected manually through surveys and interviews.

In this application, we considered a time horizon of one year. We included resources (i.e., equipment or supplies, physical infrastructure, and staff), and activities across four phases of the surgical path: (i) preoperative, (ii) operating room, (iii) post-anesthesia, and (iv) ward beds. These phases can be interpreted as a single, or

a collection of activities in the general formulation. The modeling incorporates more than 150 different resources, and surgeon is the only resource that can be shared across locations (flexible resource). Resource consumption and capacities are measured in time equivalent units. This corresponds to the multiplication of two components; the quantity (e.g., a c-arm) and the time the resource is used (e.g., 1 hour in the OR). Based on the AMC's historical data, we computed estimates of typical (average or median) resource consumptions, and we assumed they are the same across hospitals. These estimates are likely to remain valid for future surgical cases, unless significant changes in technology or surgery technique occur, in which case estimates must be updated. Estimates of capacity, on the other hand, are based on the most recent year of data. Limiting the analysis to a subset of procedures types introduces a unique challenge in capacity modeling. Specifically, resources are being used by all procedures types, including those that are not part of the model. To reconcile this discrepancy, we approximately segmented the capacity into the portion that is available for the set of studied procedure based on current utilizations, block time allocations, and volume shares. Revenue is estimated from payments data by procedure type and location. Cost is estimated from the AMC's cost reports, and includes expenses related to surgical supplies and disposable equipment, pharmacy, and other minor expenses, and we assume that it is the same across hospitals. Network capacity costs (fixed cost allocations) are excluded from these estimates, since they are basically committed well in advance. In terms of demand, we consider existing and leaked demand in the most recent year. Existing demand is estimated based on AMC's volumes and is considered as the baseline demand. Leaked demand was estimated by analyzing claims data of in-network patients that received care outside the network. These estimates of existing, and leaked demand, are used to determine the various demand bounds in the model. For a detailed description of the estimation of the model parameters see Appendix B.

3.4.2 Results

In this section, we present the main analyses and insights derived from the application of our framework. We start by describing the baseline scenario that aims to replicate the current operations at the AMC. We later use this as a benchmark for the subsequent studies of (1) the value of capacity at the AMC, and (2) the value of recovering leaked demand in the network.

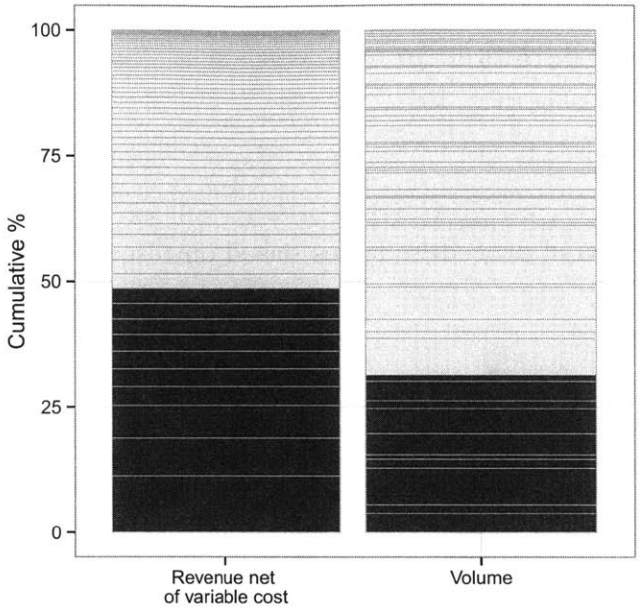
We recall that in the model we value cases based on *contribution margin* (revenue - service cost), and the objective is to maximize *revenue net of variable cost* (RNVC) (sum of volume times the contribution margin) across the set of studied procedures. We compare our results against the traditional *net contribution* (revenue - allocated cost (from cost accounting)) ranking of procedures, where allocated cost includes portions of overhead and labor costs that are allocated based on resource consumption. The net contribution ranking closely represents the way in which executives and managers traditionally value and prioritize procedures in the healthcare industry.

Baseline

In this scenario, demand is fixed according to the existing volumes at the AMC, and community hospitals capacity is excluded from the model. Thus, the entire demand is served at the AMC. To have a better understanding of the importance of the different procedure types, Figure 3-3 shows the cumulative RNVC, and cumulative volume based on the baseline volumes. We observe that there are 10 procedures types that account for 50% of the network RNVC, while their volume account for about 31% of the studied volume.

We obtain utilizations from the outputs of the optimization model; the operating room has the largest utilization, about 82%, followed by 62% and 35% in the preoperative and post-anesthesia bays, respectively. The overall surgeon time utilization is 62%. For the ward-beds capacity, we did not have a hard capacity constraint in the model, but we obtained a lower bound based on the existing volumes. The model suggests a minimum of 26.23 beds per year for the set of studied procedure types.

Figure 3-3: Baseline cumulative revenue net of variable cost and volume.



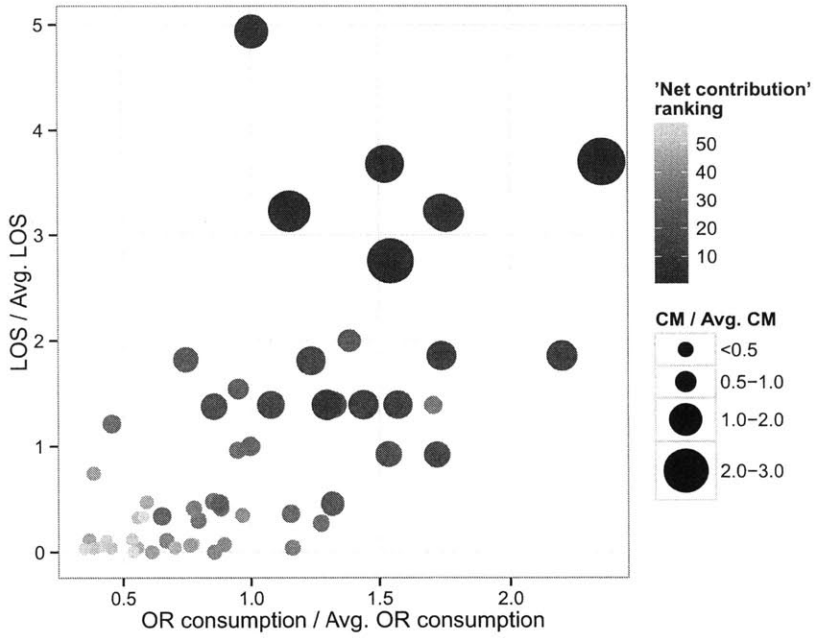
Note: Stagged bars represent procedure types. The highlighted dark area correspond to the top ten largest contributors of RNVC. They account for close to 50% of the revenue net of variable cost, and for about 30% of the volume in the set of procedure studied. Data based on FY2012.

Based on our conversations with the managers at the surgical department, these results agreed with the current operating point at the AMC. Moreover, we also concluded that the operating room utilization is already high and could not be significantly increased without a major risk of negatively affecting the daily performance (e.g., excessive overtime, delays in the schedule, etc.). In addition, ward-beds capacity is also considered as limiting resource. Therefore, we consider ward-beds and operating rooms as bottlenecks in the current AMC operations. Conversely, surgeon time utilization could still be increased. Note that surgeons demand for additional OR time at the AMC, and the issue of leaked demand, and spare surgical capacity in the community, motivated our work to start with.

To have a better understanding of the diversity of procedures types studied, Figure 3-4 shows a comparison of the different procedures types based on resource consumption and contribution. Overall, we observe that procedures types that consume more resources than the average procedure type, also tend to report higher contribution

margin (larger bubble sizes). In addition, we also observe that procedure types with high contribution margin are, although not in the same order, at the top of the net contribution ranking (i.e., dark large bubbles).

Figure 3-4: Comparison of procedure types based on resource consumption and profitability.



Note: Each point corresponds to a different procedure type. The x-axis an y-axis correspond to the ratio of operating room and ward-beds consumptions, respectively. The size of the bubbles represents the ratio of contribution margin (revenue- variable cost) over the average contribution margin in the studied set of procedures. The color of the bubbles resembles the ranking of procedure types based on net contribution margin (revenue- allocated cost), where darker is higher priority.

Finally, the contribution of surgeon's sub-specialty in terms of RNVC as well as volume is shown in Table 3.4. According to the baseline volumes, general, colorectal and breast surgery are the largest business lines. On an average per case, the sub-specialties are ranked as HPB, esophageal, colorectal, bariatric, general, endocrine, and breast.

Table 3.4: Sub-specialty contribution in baseline scenario.

Sub-specialty	% Revenue net of variable cost	% Baseline volume
Bariatric	4%	4%
Breast	12%	21%
Colorectal	27%	22%
Endocrine	9%	10%
Esophageal	2%	1%
General	36%	39%
HPB	11%	3%

Analysis 1: The value of capacity at AMC

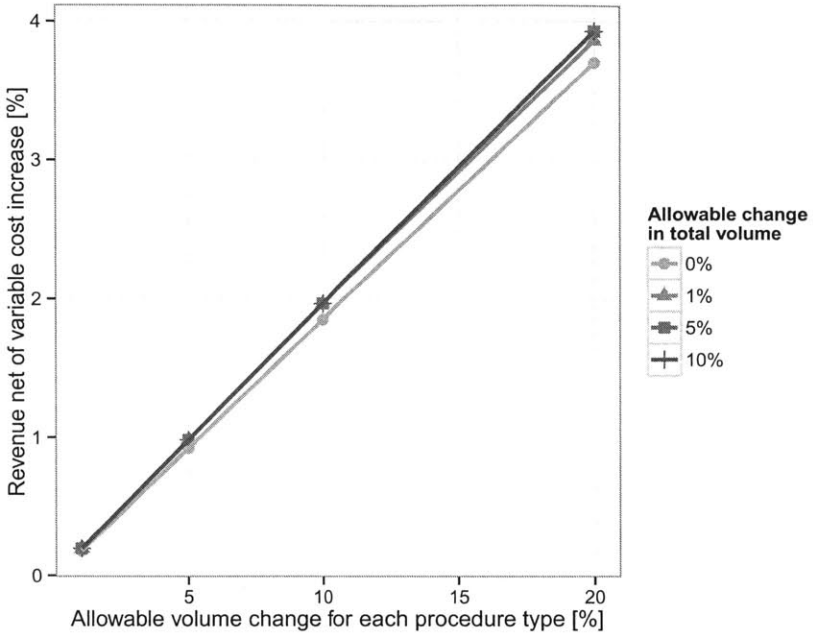
The goal of this study is to determine the potential gains that the AMC can obtain by choosing an optimal, but not significantly different, portfolio of services. The output of this analysis could be used, for example, to quantitatively identify which procedures should be grown and which should be de-emphasized at the AMC in order to improve RNVC.

As in the baseline scenario, we restrict this analysis to the AMC. We assume that capacities and utilizations are fixed according to the baseline scenario, but we allow small changes in volume within a small range around the baseline volumes. Specifically, for each individual procedure type, we allow an $\pm x\%$ volume change while overall network volume stays within $\pm y\%$ of the baseline volume. Figure 3-5 shows the potential gains in RNVC for different scenarios. By changing individual procedure types volume in up to $\pm 5\%$, a 1% increase in RNVC can be obtained. We estimate that 1% increase in RNVC will approximately[2] result in a 15% increase in hospital's bottom line.

We also analyze how the portfolio of services changes relative to the baseline volumes. We compare the outputs of the model against the traditional approach of prioritizing procedure types based on'net contribution'. This ranking resembles the way in which hospital' managers and executives strategically prioritize procedures. In Figure B-3, procedure types are ordered based on the net contribution valuation. We observe that the model suggests to decrease the volume of several procedures at end of

[2]Assumption: hospital margin is 4%, and variable cost is 40% of total cost.

Figure 3-5: Revenue net of variable cost increase at the AMC.



the ranking (less profitable procedures). This recommendation was in agreement with the expectation of the executives and managers at our partner institution. On the contrary, the decrease of the volume of procedure types 6, 7, 9, and 10 was somehow counterintuitive for them. The main difference is that the model explicitly accounts for the consumption of resources and their limited capacity. For example, procedure 6 reports a very high revenue, however, it is also a very expensive procedure in terms of usage of bottleneck resources (highest bubble in Figure 3-4). When there is limited capacity, net contribution ranking of procedures does not capture the interaction among procedures that compete for that capacity, moreover, it can lead to sub-optimal decisions. As we see in this case, instead of increasing the volume for this procedure, the optimization model reduces its volume, and use that capacity to increase the volume in several other 'less profitable' procedures.

Analysis 2: Recapturing leaked demand

In this study, we use our model to determine how surgical activities should be allocated across the network of hospitals in order to generate access that will allow the

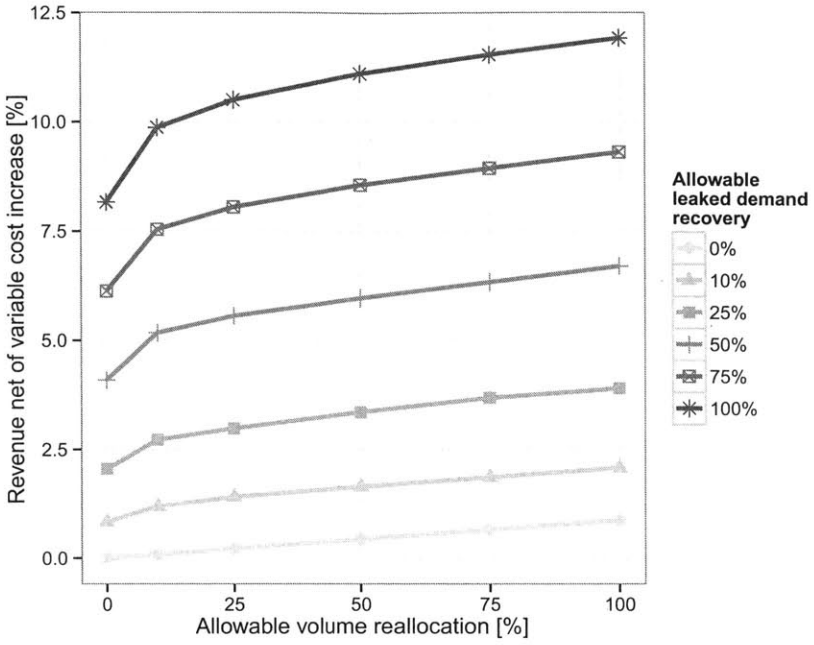
network to recapture leaked demand, while maximizing RNVC.

Contrary to the previous study, we assume that the entire network's surgical capacity is available. However, we restrict the utilization of bottleneck resources at the AMC according to the baseline scenario utilization. Preliminary runs of the model suggest that most of the resources, especially equipment, have very low utilizations, hence, we did not restrict their utilization to the baseline levels. These resources are not significantly constraining volumes at the AMC; nonetheless, coordination and scheduling of these resources will be crucial in order to ensure availability at the operational level. The capacity in the community is fixed, except for the ward-beds, for which we will explicitly report the capacity needs. In addition, we also assume that baseline demand can be reallocated to the community, and leaked demand can be recovered across the network. Thus, instead of restricting baseline demand to be performed at the AMC, we allow for reallocation to the community. We run scenarios for different reallocation and recovery levels.

In terms of RNVC increase, we estimate that, reallocation of the AMC volume across the network can result in significant gains. In Figure 3-6 we observe that by simply reallocating volumes (0% leaked demand recovery curve), RNVC can be increased by up to 1.2%. This effect becomes even more significant when we allow for leak demand recovery. Thus, if the network is able to backfill reallocated volumes by recovering leaked demand, the RNVC can be increased by up to 12% (100% volume reallocation and 100% leaked demand recovery). Additionally, we also observe that volume reallocation has decreasing marginal returns. Observe that these incremental gains correspond to much greater bottom line impact (1% increase in RNVC corresponds to approximately 15% in network's bottom line). Also, note that these gains do not require extra capacity at the AMC since spare community hospitals capacity is used instead.

In terms of utilization, we track the utilization of the AMC's resources. The operating room and ward-beds capacity utilization (see Figure B-2) decreases as more flexibility (reallocation) is permitted. Figure B-2 (a) shows the changes in the AMC operating room utilization; it decreases with reallocation but increases again with the

Figure 3-6: Revenue net of variable cost increase obtained by recovering leaked demand across the network.



recovery of leaked demand. Interestingly, ward-beds capacity seems to be the most limiting resource in the recovery of leaked demand. We observe in Figure B-2(b) that the capacity of ward-beds becomes fully utilized (reaching baseline capacity level) when more than 50% of leaked demand is recovered.

In terms of volume, this is now redistributed between the community and the AMC depending on how much reallocation and recovery is assumed in the model. For example, let us assume that reallocation is allowed for up to 10% of the baseline volumes, and that leaked demand can be recovered completely across the network, Figure B-4 summarizes the change in volumes at the AMC and in the community. Procedures are ranked from highest to lowest net contribution, remember that, this is the traditional way in which procedures are valued by executives and managers. The model suggests to move procedures at the end of this ranking to the community. This was in agreement with our partner's expectations. However, the model also suggests to move procedures at the top of the ranking (procedures types 2, 3, 6, and 7). This was surprising and somehow counterintuitive for the executives and managers. The reason is that these procedure types consume a large quantity of bottleneck resources

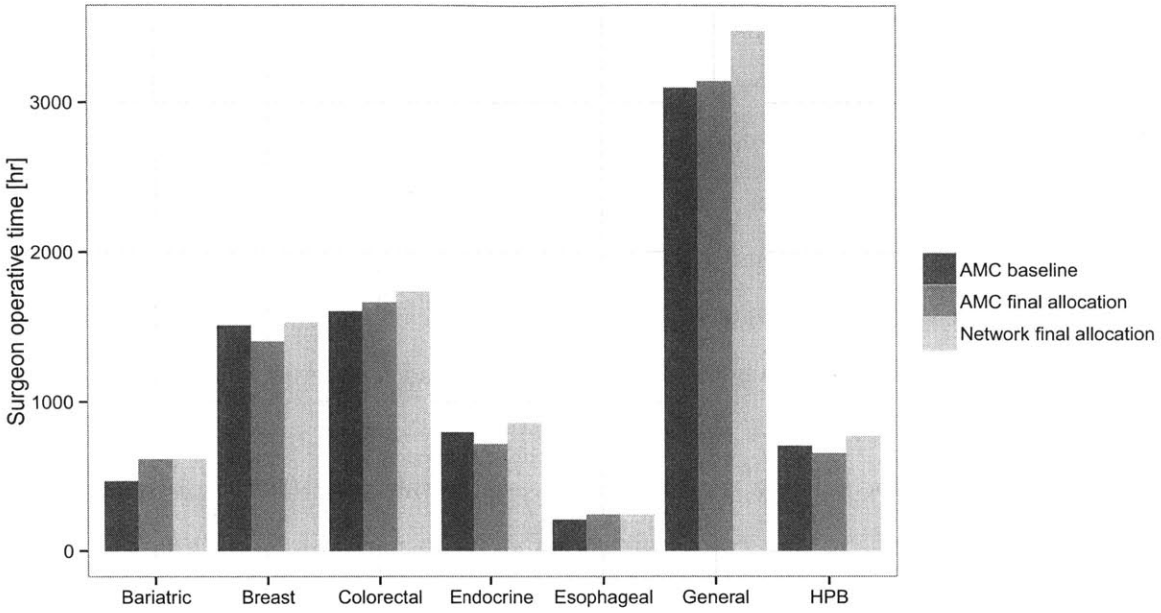
(they correspond to the largest and darkest bubbles in the upper middle area in Figure 3-4). Thus, by offloading cases of these procedures to the community, the model is able to free capacity and backfill it with cases of procedures that, under the traditional view, are seemingly less profitable, which ultimately results in better value out of the entire network capacity.Another interesting output of the model is the allocation of surgeons' time across the network. In Figure 3-7, we show an example of how the surgeons' operative time allocation changes in the AMC and across the network assuming a 10% volume reallocation, and full leaked demand recovery. Even in the case where all leaked demand is recovered, the overall surgeon time utilization is below 70% (in the baseline the utilization is 62%). In general, we observe that sub-specialties, such as breast, endocrine and HPB transfer operative time from the AMC to the community, while the other sub-specialties increase or maintain their presence at the AMC. According to the baseline volumes, breast and endocrine have the lowest average contribution per patient, and HPB has the highest. This latter recommendation was, again, somehow counterintuitive for the managers and executives, but it demonstrated them how traditional approaches might fail to capture complex interactions in a network environment. On the other hand, at this point in time and independently, of this study, our partner started to offloading some breast and endocrine procedures to a recently acquired clinic, our analysis was in agreement with such strategy.

3.5 Conclusion

In this paper, we proposed a general framework to support strategic decision making in healthcare delivery network. Specifically, we developed a linear optimization model akin to revenue management approaches. The model captures relevant operational constraints, thus ensuring that a particular mix of surgical cases can actually be performed in practice. Using this model, healthcare networks can budget capacity and resources to specific services, such that network's welfare objectives are maximized.

Our model overcomes some of the main difficulties of common approaches. Heuris-

Figure 3-7: Example of total surgeons' operative time by sub-specialty at the AMC, and across the network.



Note: The final allocation assumes up to 10% reallocation, and full leaked demand recovery. The 'AMC baseline' scenario corresponds to baseline volumes performed at AMC. The 'AMC final allocation' is the optimal total operative time at the AMC and the 'Network final allocation' is the total operative time across the entire network (AMC + COM), both assuming 10% reallocation and full leaked demand recovery.

tics often fail to properly capture the opportunity cost of alternative actions; this is particularly true when the space of alternatives is large. Moreover, ad-hoc approaches often include fixed cost of capacity, which is arbitrarily assigned to procedures and activities, and use this to support various strategic decisions. This approach incorrectly favors (diminishes) procedures which superficially appear to have low (high) overhead; our approach strictly model the consumption of resources and only include direct and variable cost of providing the service.

We presented an application of the model to a network of hospitals that consists of two community hospitals and one academic medical center. The goal was to determine how to better use the spare surgical capacity in the community in order to recapture leaked demand while maximizing RNVC across the network. We conducted several analyses that demonstrated how the model outputs differ from traditional practices,

specifically, the way in which procedure types are prioritized. The results revealed significant and practical managerial insights to the executives and managers in our partner network. Our approach has the potential to radically transform the way in which healthcare networks understand cost and allocate resources in practice.

Chapter 4

The Nature and Sources of Variability in Pediatrics Surgical Case Duration

4.1 Introduction

Understanding and predicting surgery duration has motivated extensive research in operating room (OR) management, statistics, and operations research. Despite this, uncertainty in case duration continues to affect efficiency and productivity by causing long waiting times, misallocation of resources, and difficulties with team coordination [27]. In addition to the obvious operational implications, there are also research implications since variability can distort retrospective periopeative outcome studies that use actual case duration, rather than scheduled surgery time, as an independent variable in logistic regression models [23].

A large body of work aims to support various decisions in the OR [27] by providing accurate estimates of surgical time medians, bounds, and remaining surgery time [26, 28] for individual cases or lists of cases. These decisions may include whether to perform a specific case, how to sequence cases in a specific suite [29], how to swap cases between suites [81], how to update estimates of remaining surgery time [26], or how to address specific issues such as add-on cases [105], cancellations, and staffing necessary to complete a list of cases [46]. Research has sought to identify factors associated with surgery duration, hoping to control for variability factors and

support better decisions [33, 79]. Additionally, knowing which factors correlate with surgery duration is also useful in retrospective outcomes studies.

Prediction methods are largely developed for adult surgery, but there is little corresponding work in pediatrics. Some studies have highlighted the importance of distinguishing pediatric from adult institutions [56, 96], but attempts have been limited to single, simple procedures [81]. Moreover, there is limited quantitative information of any kind by which payers and policy makers can compare the complexity and diversity of pediatric procedures to adults.

Unlike adult hospitals, pediatric hospitals have the burden of managing extremes of variability that arise from an especially unpredictable patient population [56]. Wide ranges in patient age, size, weight, and developmental level are superimposed upon an even wider range of pathology. We hypothesized that the predictive power of commonly used patient and procedure factors is less substantial in pediatrics. To test this hypothesis, we explored the nature of case time variability within a wide range of surgeries carried out over several years in a large, urban, academic pediatric hospital. Our analysis introduces two original pieces. We begin by offering the first quantitative description of case-time variability across a wide range of pediatric surgeries. We then use a regression tree method [84] to test factors found to influence case time among adults, and evaluate the quality of those predictors by comparing improvement against the standard surgery time allocation method in place.

4.2 Methods

This research was conducted under the Quality Improvement initiative at the Department of Anesthesiology, Perioperative, and Pain Medicine of Boston Children's Hospital.

4.2.1 Data description and current prediction method

Data description

We examined all scheduled surgeries over three years (43,000 elective non-cardiac procedures) using information obtained from the administrative database of a large, urban, academic pediatric hospital. Procedures were indexed by the internal coding system that forms the basis of managerial decisions in our institution [100]. Overall, records represented more than 1,500 different in-house procedure types, performed in 21 ORs by more than 300 surgeons from 16 surgical departments.

We define case time as the time from patient entry-to-exit from the OR ("wheels in" to "wheels out"). For each procedure type, extreme values (most likely due to data entry errors) were filtered through logarithmic transformation [87] and removal of all observations falling more than three standard deviations from the mean. In this process, fewer than 1% of data was lost, leaving 42,505 records. Additionally, non-identifying patient characteristics including age, weight, ASA (American Society of Anesthesiologists physical status), and ICU request indicator were available for 65% records. The volume was further split into train (FY2008 and FY2009) and test data (FY2010). Train data is used to build models while test data is used to evaluate the performance of the methods.

Standard surgery time prediction method

In this hospital, and many others, the prediction of case time is based on a surgeon-procedure specific moving average with rolling horizon of 5 observations. After removing extreme values, this method assigns as predicted time the average of the five most recent surgeries performed by each surgeon. Extreme values are identified

according to the previous definition, and observations were collected until the most recent 5 non-extreme values are obtained. When a surgeon has performed fewer than 5 surgeries for a given procedure, predictions are based on historical average.

At our institution, the rolling prediction can be modified by surgeon requests for more or less time. Although such adjustments have improved estimates in some settings [33, 86, 100] changes to the rolling average estimates are restricted at our institution due to experience with large underestimation bias. We consider this as the "standard" prediction in the current system and use it for later comparison to an alternative prediction method.

4.2.2 Analysis 1: Case time variability and performance of the standard method

We study 249 procedure types for which 30 or more observations exist. This encompassed 33,302 scheduled procedures (78.4% elective volume). For each procedure, we computed bootstrap point estimates [80] for the mean, median, standard deviation (SD), minimum/maximum, and quantiles of case time. Shapiro-Wilk test of normality revealed that most do not distribute normally. The null hypothesis of normality was rejected for 209/249 procedure types while log-normality was only rejected for 83/249 procedure types [88, 87]. Given this, and that sample size varies significantly among procedures, we used a bias-corrected and accelerated bootstrap method [80] with 10,000 re-samples to compute statistics and confidence intervals.

To understand the diversity of these procedures, we compared the duration and variability statistics using plots and linear regression. To measure the performance of the standard prediction method, we compared the Mean Absolute Error ($MAE = \frac{1}{n}\sum_i |e_i|$, where $e_i = \hat{y}_i - y_i$ is the error of the prediction $\hat{y}_i$, and y_i is the actual duration of observation i), and the Mean Absolute Percentage Error ($MAPE = 100\%\frac{1}{n}\sum_i \left|\frac{e_i}{y_i}\right|$) of the test sample predictions for 195/249 procedure types. We limited this analysis to 195 procedure types (72% elective volume) so that train data contained at least 20 observations and test data at least 10.

4.2.3 Analysis 2: Prediction of surgery duration based on commonly used factors

To evaluate predictive factors, we applied conditional inference regression trees (RT) to individual procedure types. This is a non-parametric learning technique whereby trees are formed based on factors within the modeling dataset. Based on train data, trees are constructed iteratively, such that rules based on independent factors are selected to provide the best split along a dependent variable (case time). When significant relationships exist between the independent and dependent variables, data is split into two groups that maximize differences in the dependent variable. The branching criterion is such that the variable with stronger association to case time is chosen. The stronger associated variable is identified by constructing a permutation test, for the independence between case time and the prediction variables to then select the variable with the minimum Bonferrroni adjusted p-value [48, 73]. Once the splitting variable is chosen, a binary split occurs that is defined to make the resulting daughter nodes more dissimilar. Hence the splitting criterion is given by the splitting value that maximizes the test statistics measuring the discrepancy between the two daughter nodes. This process is repeated recursively and the stopping criterion for branching is met whenever the global null hypothesis of independence between the dependent variable and the predictive variables cannot be rejected with confidence level 5%. This statistical approach ensures that no pruning or cross validation is needed (see [48]). The desired end result is the understanding of the combination of variables that have strong association with case time for each specific procedure.

The final buckets in the tree contain train observations that are similar to each other and the mean of their case time can be used to predict the duration for out-of-sample surgeries. Thus, surgery time for a new patient is allocated by classifying the patient into one of the buckets and assigning the mean of the bucket. Classification is based on the patients specific factors following the tree's splitting criteria from the root to the leaves.

The independent factors included surgeon, ICU bed request, ASA status, patient

age, and weight. These parameters are easily accessible and have been offered as predictive elsewhere [33, 83, 89]. Complete information was available for 27,456 elective surgeries. To obtain sample sizes containing no fewer than 20 train examples and 10 or more test samples, we further limited our analysis to 120 procedure types. To insure that train and test samples come from the same distribution, we performed a two-sample Kolmogorov-Smirnov (K-S) test for all 120 procedures. High p-values returned for most procedures but null hypothesis was rejected for 12 procedure types. As a result, this portion of the analysis was confined to 108 procedure types (55% elective volume).

To evaluate the accuracy gained by incorporating these features, we compared the prediction errors of the RT predictions against the standard method using test samples. Notice that this is the most naive comparison possible since the standard method is just a simple moving average of surgeon's most recent cases.

a) We compared the distribution of prediction errors using the two-sample K-S test for each procedure type for which the RT finds correlation between case time and the studied factors (i.e., procedure types for which the RT splits).

b) Secondly, we compared the global metrics MAE and MAPE, and the 10%-quantile of prediction errors for procedures for which the RT splits using linear regression. Each procedure type corresponds to an observation and we regress the RT metric on the respective standard method metric. The slope of the regression line is interpreted as the average prediction improvement of the RT over the standard method for procedures included in the regression. For example, if the slope is less than one, the RT provides more accurate predictions than the standard method. Finally, we perform this same analysis leaving the 10% longest surgeries for each procedure type aside in order to quantify the performance of the RT when extremely long cases are excluded.

4.3 Results

4.3.1 Analysis 1: Case time variability and performance of the standard method

Pediatric surgeries are widely variable and this variability does not cluster

Across all 249 procedure types, the SD of case time was 30% (25.8, 34.7) of the median as defined by the regression slope (Figure C-1). The notation $x(\underline{x}, \bar{x})$ corresponds to the point estimates and the 95% confidence intervals (CI). As previously observed in Strum et al. [89], we found that relative variability, as measured by coefficient of variation (CV, Figure C-2), was greater for shorter than for longer procedures and that some procedures (those above the regression line in Figure C-1) were significantly more variable than others. In addition, Figure C-1 illustrates that most procedures tend to cluster around the trend line but many are outside the 95% CI. The range of variability was high, with procedures spanning wide ranges and others being fairly narrow. For example, *Open Gastrostomy, General Surgery* had a median duration of 172(147, 213) minutes and SD of 140(91, 206) minutes. Conversely, *Lefort I, Plastic* procedures were performed with a median duration of 315(285, 331) minutes and SD of 115(96,142) minutes. Unfortunately, procedures with outlying variability did not cluster by type, location, or specialty of surgery, making it difficult to identify them prospectively.

Performance of the standard prediction method

In our population (195/249 procedures), the standard method performs poorly. In Figure C-3(a), we observe that longer procedures present larger absolute deviations from scheduled durations, i.e., large MAE. Overall, the current method results in average deviations of 19.6(15.6, 24.8)% from the median case time (slope of the regression line in Figure C-3(a)). In relative terms, MAPE (Figure C-3(b)) was greater than 40% for 43/195 procedure types, that is, the predicted duration was inaccurate on average in more than 40% for these 43 procedure types. This large relative de-

viations are more common among shorter surgeries. Additionally, the MAPE was between 20 and 40% for half of the procedures, and was below 20% in only 43/195. Disappointingly, procedures in the final group accounted for only ~13% of the total volume.

4.3.2 Analysis 2: Prediction of surgery duration based on commonly used factors

We used surgeon identity, ASA status, patient age and weight, together with the ICU bed request factor, to build RTs for each procedure type. As one example, Figure C-4 shows the RT for the procedure *Unilateral Orchidopexy*, built using 265 train observations. The overall mean ± SD of case time was 93.25±28.19 minutes. The tree split the initial bulk of observations based on surgeon identity, ASA, and age factors.

Overall, we found significant splits, along single factors or combinations, in only 39/108 procedures (36%). Most notably, surgeon identity correlated with case time in only 27/108 procedure types (25%) and, for 9 of these, surgeon identity was not the only relevant variable. ICU bed request was found to be relevant for only 3 procedures (ICU indicator is meaningful for 18 procedures), ASA status for 8, and patient age or weight for 15 procedures. The Table C.1 presents a summary of the 39 procedure types for which the RT split, as well as the order in which factors were used in the splitting process. For the 39 procedure types where case time and factors did correlate, we measured the predictive performance of the RT:

a) Compare distribution of the prediction errors in test samples of the RT and the standard method

    i) Procedure types for which the RT split based on surgeon identity (27 procedure types). The prediction error distributions were significantly different for only 4/27 procedures (procedures 2, 3, 12, and 28 in Table C.1), as determined by K-S test at the 5% significance level. Furthermore, we observe that the RT reduced overestimation error but no significant improvement

was observed in the underestimation error (i.e., actual time is larger than predicted).

ii) Procedure types for which the RT split on factors other than surgeon identity (12 procedures). As determined by the K-S test, there was no significant difference between the prediction errors of the methods except for two procedures, 35 and 43 in Table C.1. In the former, RT reduces the median prediction error, while in the later the median increases but underestimation errors are reduced. Therefore, we found no significant evidence of positive effect in prediction by controlling for surgeon identity when using the standard method for this set of 12 procedures.

b) Compare the MAE and MAPE of the regression tree and the standard method for 37/39 procedure types (procedure types 76 and 80 were excluded because of outliers behavior). We regressed the MAEs of the RT on the MAEs of the standard method, the slope was 0.82 (R-squared 0.78). When using MAPE instead, the slope was 0.67 (R-squared 0.77). The performance is affected by extremely long surgeries that will record large prediction errors and skew the metrics. To investigate this, we compared the MAEs and MAPEs when extremely long cases are left aside. We found that by leaving out the 10% longest cases for each procedure type, the MAEs of the RT become even smaller than the standard method (regression slope decreases up to 0.67 (R-squared 0.53)). The MAPEs also improved; the slope reduces to 0.61 (R-squared 0.67). Finally, we compare the underestimation error by comparing the 10%-quantile of the prediction errors of the two methods. We regressed the 10%-quantile underestimation error of the RT over the standard method, the slope of the regression line decreases from 1.03 (R-squared 0.81) to 0.62 (R-squared 0.44) when the 10% longest surgeries are excluded.

4.4 Discussion

Imperfect surgical scheduling results in idle time, overtime, and substantially increased cost. Large variability from allocated surgical times and the inability to anticipate such deviations present obvious challenges for operating room managers. We found that a method of allocating surgery time based only on procedure type and surgeon identity performs poorly in our pediatric population. Large deviations from scheduled durations were observed as indicated by large MAEs and MAPEs (>20% for 152/195, and >40% for 43/195 procedure types). To improve surgical scheduling, we then studied the utility of commonly used patient and procedure factors in predicting pediatric case times. Using regression trees, we determined that surgeon identity, ICU bed request, ASA status, patient age and weight, alone or together in any combination, had little association with case time in our large and diverse population. Largely, the failed associations arose from high variability within each procedure type and the long-tail behavior of case times. Future research in pediatric OR operations should seek to prospectively identify cases that have extreme behavior. Until better predictors can be identified, scheduling inaccuracy will persist.

Our findings differ substantially from similar studies of adult surgery. One potential explanation is that pediatric surgery is simply more variable than adult surgery. We considered more than 40,000 pediatric surgeries and analyzed the most common 249 procedure types. Although no global metric can accurately capture all aspects of case time variability, we can begin by comparing our calculations to adult reports. Spangler et al. [83] report a CV for 574 procedure types for which there were at least 27 observations. There, 98.4% had CV less than 0.76. Strum et al. [89] studied 40 CPT codes, finding that 92.5% of procedure types had empirical CV's below 0.5. Dexter et al. [25] estimated the CV for 354 procedure types, finding 90% to have CV below 0.5 and 99.4% below 1.0. In our pediatric population, considerably more of the most common procedures (53/249, or nearly one quarter) had CV above 50%. This fact, together with the observation that 160/249 procedure types did not reject for log-normality (i.e. had a right heavy-tail distribution) suggest that high

variability and unanticipated long surgeries are more common in pediatric than adult populations.

Many investigators have studied case time variability in adult surgery. While an exhaustive review of that literature is beyond our scope here, many studies [24, 33, 89, 86, 85] have sought single variables correlating with case time and concluded that surgeon identity is one of the most important sources of variability. Some investigators, such as Silber et al. [79], have considered more complex relationships and sought to describe combinations of patient characteristics that associate with case time. Dexter et al. [22], observed that procedure classification is important issue and that uncommon procedures drive variability. One elegant approach is a Bayesian method to create bounds on the estimates of uncommon procedures' time [28].

Despite all of the work in adult centers, pediatric case time variability is specifically addressed in only a single study of endoscopy procedures [81]. We therefore used regression trees to search for factors that associate with case time variability and that might prove useful in scheduling. This powerful machine learning technique can capture non-linear relations and rapidly test the predictive value of many variables, alone or in combination, across large datasets. To our knowledge, this is the first application of this technique to surgical scheduling. Surprisingly, none of the variables previously associated with case time in adults were generally correlated with case time in our pediatric population. In particular, although surgeon identity has frequently been shown to correlate with case time in adult surgery [86, 89, 24], we found only very limited association in some procedures and no association in the vast majority. In fact, for procedures where association with surgeon identity did exist, the tree always split along groups of surgeons and never on individuals. Importantly, these groupings also failed to extend to other procedures. Our familiarity with the data suggests that variability arising from the surgeon is inconsistent across procedures here and is generally overwhelmed by other factors.

Our findings have four practical implications for pediatric surgery. First, managerially, they suggest that attentions focused on individual surgeons may be largely misplaced. Second, in scheduling, they show that algorithms using surgeon-specific

case times are unnecessary. Third, for researchers, they suggest that retrospective studies may be simplified and sample size issues resolved by pooling data from multiple surgeons and reducing the number of combinations (i.e. surgeon-procedure) in statistical models. Finally, we show that case-time prediction accuracy can be significantly improved by prospectively identifying extreme behavior cases. Although the regression tree method was unable to isolate such cases using covariates studied here, it is a powerful tool for screening additional variables and should prove useful in future efforts. Until better predictors of pediatric case time are identified, daily OR management in pediatric centers will likely require more overtime, capacity buffers, and schedule flexibility than in adult centers.

4.4.1 Limitations

Our experience is that of a single large, urban, academic center and may not be generalizable to other institutions. Rare and physiologically complex procedures are common in pediatric hospitals [56] and we believe that this accounts for much of the variability. To the extent that this is true, our findings should be applicable to institutions serving similar populations. At the same time, process and personnel variables unique to our institution may have increased or decreased differences that naturally arise from the patient and procedure. If so, the analytical techniques employed are broadly applicable and should prove useful in identifying such variables. This is an active area of our present research.

Our work also involves the general limitations described in similar studies [33]. Most importantly, statistical analysis of historical experience requires sufficient numbers of cases with consistent and accurate naming of procedures. Here, we confined our analysis to scheduled procedures where sufficient numbers were available to yield statistical confidence. We used internal procedure names rather than names from standardized systems (e.g. CPT codes) since this is the basis of operational decisions at our institution. While we believe this was warranted, it is possible that naming conventions led to aggregation or disaggregation of cases that could influence results [22].

Chapter 5

Concluding Remarks

In this thesis, we studied strategic and operational challenges arising at three different levels - market, system (i.e., network), and organization - in the healthcare industry. Using a variety of techniques, such as optimization, game theory, and machine learning, we developed novel frameworks that allow us to better understand cost and resource allocations for informing strategic decision making in healthcare settings.

The U.S. healthcare industry is undergoing a massive transformation process. The policy and regulation changes introduced by the recently enacted healthcare reform have resulted in the formation of large healthcare delivery systems. Moreover, the change in the reimbursement system and the focus on quality of outcomes has resulted in a whole new set of incentives which is driving the delivery of care towards a more patient centric one. Unfortunately, traditional practices in the industry are often inadequate, or outdated to face the new challenges. The transformation of the industry requires understanding markets, systems, and organizations trade-offs, and it needs now, more than ever before, novel approaches to address contemporary operational and strategic challenges in the post-reform era. We believe that operations management frameworks, like the ones developed in this thesis, can significantly contribute to this purpose and improve decision making in healthcare settings.

Appendix A

Proofs Chapter 2

A.1 Proof of Lemma 2.1

Proof. We consider discount and premium parameters $\alpha \in [0,1]$ and $\gamma \geq 0$. We assume that the SP is loss averse and model his utility using a piecewise linear function. That is, the utility parameter $\nu_{SP} > 1$ discounts losses heavier than equivalent gains (which has discount factor 1). In this Lemma we consider a general new contract $C^{new}(D)$ with non-negative and non-increasing marginal cost of service. In addition, for a contract to be feasible, it must satisfy the risk reduction and extra-payment constraints. Thus, under this conditions, the total payment curve induced by a feasible new contract must cross the payment curve of the standard contract in exactly one point, d^*.

Let us consider a feasible contract that does not necessarily satisfy risk reduction and extra-payment constraints with equality. Thus, we consider $K_1 > 0$ and $K_2 > 0$ such that

$$\begin{aligned} \mathrm{E}[C^{new}(D) - C^{std}(D)] =& \gamma \mathrm{E}[C^{std}(D)] - K_1 \\ \mathrm{E}[C^{new}(D) - C^{std}(D) | D > d_\beta] =& -\alpha \mathrm{E}[C^{std}(D) | D > d_\beta] - K_2 \end{aligned} \tag{A.1}$$

The new contract induces demand break even point d^*, and we compute the resulting expected utility respect to it. The new contract induces demand break even

point d^*, and we compute the resulting expected utility respect to it.

- **Case 1:** $d^* \leq d_\beta$

  Using the equalities in (A.1),

$$\begin{aligned}
&\mathrm{E}[C^{new}(D) - C^{std}(D) | D \leq d^*] F(d^*) = \\
&\mathrm{E}[C^{new}(D) - C^{std}(D) | D \leq d_\beta] \beta \\
&- \mathrm{E}[C^{new}(D) - C^{std}(D) | d^* < D \leq d_\beta] (\beta - F(d^*)) \\
&= \gamma \mathrm{E}[C^{std}(D)] - K_1 - \mathrm{E}[C^{new}(D) - C^{std}(D) | D > d_\beta](1 - \beta) \\
&- \mathrm{E}[C^{new}(D) - C^{std}(D) | d^* < D \leq d_\beta] (\beta - F(d^*))
\end{aligned}$$

$$\begin{aligned}
&\mathrm{E}[C^{new}(D) - C^{std}(D) | D > d^*](1 - F(d^*)) = \\
&(-\alpha \mathrm{E}[C^{std}(D) | D > d_\beta] - K_2)(1 - \beta) \\
&+ \mathrm{E}[C^{new}(D) - C^{std}(D) | d^* < D \leq d_\beta] (\beta - F(d^*))
\end{aligned}$$

  The expected utility induced by the new contract corresponds to

$$\begin{aligned}
&\mathrm{E}[U_{SP}(C^{new}(D) - C^{std}(D))] = \mathrm{E}[C^{new}(D) - C^{std}(D) | D \leq d^*] F(d^*) \\
&- \nu_{SP} \mathrm{E}[C^{std}(D) - C^{new}(D) | D > d^*](1 - F(d^*)) \\
&= \underbrace{\gamma \mathrm{E}[C^{std}(D)] - (\nu_{SP} - 1) \alpha \mathrm{E}[C^{std}(D) | D > d_\beta](1 - \beta)}_{constant} \\
&- K_1 - (\nu_{SP} - 1) K_2 (1 - \beta) \\
&+ (\nu_{SP} - 1) \underbrace{\mathrm{E}[C^{new}(D) - C^{std}(D) | d^* < D \leq d_\beta]}_{\leq 0} (\beta - F(d^*))
\end{aligned} \tag{A.2}$$

- **Case 2:** $d^* > d_\beta$

Similarly to the previous case,

$$\begin{aligned}
&\mathrm{E}[C^{new}(D) - C^{std}(D) | D \leq d^*] F(d^*) = \\
&\gamma \mathrm{E}[C^{std}(D)] - K_1 - \mathrm{E}[C^{new}(D) - C^{std}(D) | D > d_\beta](1-\beta) \\
&+ \mathrm{E}[C^{new}(D) - C^{std}(D) | d_\beta < D \leq d^*](F(d^*) - \beta)
\end{aligned}$$

$$\begin{aligned}
&\mathrm{E}[C^{new}(D) - C^{std}(D) | D > d^*](1 - F(d^*)) = \\
&(-\alpha \mathrm{E}[C^{std}(D) | D > d_\beta] - K_2)(1-\beta) \\
&- \mathrm{E}[C^{new}(D) - C^{std}(D) | d_\beta < D \leq d^*](F(d^*) - \beta)
\end{aligned}$$

The expected utility induced by the new contract corresponds to

$$\begin{aligned}
&\mathrm{E}[U_{SP}(C^{new}(D) - C^{std}(D))] = \\
&\underbrace{\gamma \mathrm{E}[C^{std}(D)] - (\nu_{SP} - 1)\alpha \mathrm{E}[C^{std}(D) | D > d_\beta](1-\beta)}_{constant} \\
&- K_1 - (\nu_{SP} - 1) K_2 (1-\beta) \\
&- (\nu_{SP} - 1) \underbrace{\mathrm{E}[C^{new}(D) - C^{std}(D) | d_\beta < D \leq d^*]}_{>0} (F(d^*) - \beta)
\end{aligned}$$

With this on hand, we can show the properties of the utility maximizing contract;

(i) The risk reduction and extra-payment constraints are tight.

Note that the expected utility is, in both cases, decreased due to the constraints not being tight, that is, $K_1 > 0$ and $K_2 > 0$. Therefore, any utility maximizing contract must be such that, both, the risk reduction and extra-payments constraints are met with equality, that is $K_1 = K_2 = 0$.

(ii) The demand break even point $d^* = d_\beta$.

Note that the expected utility is, in both cases, decreased due to $d^* \neq d_\beta$. In Case 1, if the demand break even point coincides with the threshold d_β, the term $(\nu_{SP} - 1)\mathrm{E}[C^{new}(D) - C^{std}(D) | d^* < D \leq d_\beta](\beta - F(d^*))$ is zero, and the

expected utility (A.2) increases. Hence, any utility maximizing contract must induce demand break even point equals to the demand threshold d_β.

Therefore, a contract is utility maximizing if and only if (i) and (ii) are both true. Moreover, the maximum expected utility corresponds to

$$EU^*_{SP}(\alpha,\gamma) = \gamma \mathrm{E}[C^{std}(D)] - (\nu_{SP}-1)\alpha \mathrm{E}[C^{std}(D)|D > d_\beta](1-\beta)$$

□

A.2 Proof of Lemma 2.2

Proof. Consider a two-price incremental discount contract. By setting the risk reduction and the extra-payment constraints to equality, we can solve the system of equations to find P_1 and P_2 as a function of the breakpoint b, and the discount and premium parameters $\alpha \in [0,1]$ and $\gamma \geq 0$. To simplify notation we use $\overline{D}_d = \mathrm{E}[D|D > d]$ and $\underline{D}_d = \mathrm{E}[D \leq d]$.

The extra-payment constraint corresponds to $\mathrm{E}[C^{new}(D)] = P_1\bar{D} - (P_1 - P_2)(\overline{D}_b - b)\overline{F}(b) = (1+\gamma)P_0\bar{D}$. Thus, solving for P_1

$$P_1(b,\alpha,\gamma) = \frac{(1+\gamma)P_0\bar{D} - P_2(b,\alpha,\gamma)(\overline{D}_b - b)\overline{F}(b)}{\bar{D} - (\overline{D}_b - b)\overline{F}(b)} \tag{A.3}$$

For the second price we solve depending on whether the breakpoint is above or below the demand threshold d_β

Case 1: $b \leq d_\beta$

The left hand side of the risk reduction constraint can be expressed as $\mathrm{E}[C^{new}(D)|D > d_\beta] = P_1 b + P_2(\overline{D}_{d_\beta} - b)$.

Case 2: $b > d_\beta$

The left hand side of the risk reduction constraint can be written as

$$\mathrm{E}[C^{new}(D)|D > d_\beta] = P_1\Big(\overline{D}_{d_\beta} - \frac{\overline{D}_b\,\overline{F}(b)}{1-\beta} + \frac{b\,\overline{F}(b)}{1-\beta}\Big) + P_2\Big(\frac{\overline{D}_b\,\overline{F}(b)}{1-\beta} - \frac{b\,\overline{F}(b)}{1-\beta}\Big)$$

Then, solving for P_2 in each case,

$$P_2(b,\alpha,\gamma) = \begin{cases} P_0 \frac{(1-\alpha)\overline{D}_{d_\beta}\left(\bar{D}-(\overline{D}_b - b)\overline{F}(b)\right)-(1+\gamma)\bar{D}b}{\overline{D}_{d_\beta}\left(\bar{D}-(\overline{D}_b-b)\overline{F}(b)\right)-\bar{D}b} & b \leq d_\beta \\ P_0 \frac{(1-\alpha)\overline{D}_{d_\beta}(1-\beta)\left(\bar{D}-(\overline{D}_b - b)\overline{F}(b)\right)-(1+\gamma)\bar{D}\left(\overline{D}_{d_\beta}(1-\beta)-(\overline{D}_b-b)\overline{F}(b)\right)}{(\overline{D}_b-b)\overline{F}(b)(\bar{D}-\overline{D}_{d_\beta}(1-\beta))} & b > d_\beta \end{cases} \tag{A.4}$$

Now, we show that the resulting prices are continuous and non-increasing in the breakpoint and they satisfy $P_2(b,\alpha,\gamma) \leq P_0 \leq P_1(b,\alpha,\gamma)$. To show continuity we just need to evaluate the second price from both cases, $b \leq d_\beta$ and $b > d_\beta$, as $b \to d_\beta$. Using the definition of $P_2(b,\alpha,\gamma)$ for $b \leq d_\beta$,

$$\begin{aligned} P_2(d_\beta,\alpha,\gamma) &= P_0 \frac{(1-\alpha)\overline{D}_{d_\beta}\left(\bar{D}-(\overline{D}_{d_\beta}-d_\beta)(1-\beta)\right)-(1+\gamma)\bar{D}d_\beta}{(\bar{D}-\overline{D}_{d_\beta}(1-\beta))(\overline{D}_{d_\beta}-d_\beta)} \\ &= P_2(b \to d_\beta,\alpha,\gamma) \end{aligned}$$

The last equality corresponds to the closed form of $P_2(b,\alpha,\gamma)$ in the case $b > d_\beta$ and we take $b \to d_\beta$. Hence, the second price is a continuous function of the breakpoint, thus, the first price is also continuous in the breakpoint. Next, we show the first order behavior of the prices.

- $P_1(b,\alpha,\gamma)$ is non-increasing and greater than P_0

Case 1: $b \leq d_\beta$

The first price is strictly decreasing in the breakpoint in this interval. We first replace the closed form solution of $P_2(b,\alpha,\gamma)$ into the closed from expression for P_1 (equation (A.3)),

$$P_1(b,\alpha,\gamma) = P_0\left(1+\frac{\gamma\bar{D}(\overline{D}_{d_\beta}-b)+\alpha\overline{D}_{d_\beta}(\overline{D}_b-b)\overline{F}(b)}{\overline{D}_{d_\beta}(\bar{D}-(\overline{D}_b-b)\overline{F}(b))-\bar{D}b}\right) > P_0$$

$$P_1'(b,\alpha,\gamma) = -\frac{(\alpha+\gamma)P_0\bar{D}\overline{D}_{d_\beta}\overline{F}(b)(\overline{D}_{d_\beta}-\overline{D}_b)}{\left(\overline{D}_{d_\beta}(\bar{D}-(\overline{D}_b-b)\overline{F}(b))-\bar{D}b\right)^2} < 0$$

Case 2: $b > d_\beta$

Similarly, we use the closed from solution of $P_2(b,\alpha,\gamma)$ for $b > d_\beta$ to find $P_1(b,\alpha,\gamma) =$

$P_0\left(1+\frac{\gamma\bar{D}+\alpha\overline{D}_{d_\beta}(1-\beta)}{\bar{D}-\overline{D}_{d_\beta}(1-\beta)}\right) > P_0$, constant in b. Therefore, the first price is non-increasing, and its minimum value corresponds to the last expression which is greater than P_0.

- $P_2(b,\alpha,\gamma)$ **is non-increasing and smaller than** P_0

Case 1: $b \leq d_\beta$

$$P_2^{'}(b,\alpha,\gamma) = -\frac{P_0(\gamma+\alpha)\bar{D}\overline{D}_{d_\beta}\underline{D}_b F(b)}{\left(\overline{D}_{d_\beta}(\bar{D}-(\overline{D}_b-b)\overline{F}(b))-\bar{D}b\right)^2} < 0$$

Case 2: $b > d_\beta$

$$P_2^{'}(b,\alpha,\gamma) = -\frac{P_0(\gamma+\alpha)\bar{D}\overline{D}_{d_\beta}(1-\beta)}{(\overline{D}_b-b)^2\overline{F}(b)(\bar{D}-\overline{D}_{d_\beta}(1-\beta))} < 0$$

Therefore, $P_2(b,\alpha,\gamma)$ is non-increasing, and
$\lim_{b\to 0} P_2(b,\alpha,\gamma) = P_0\left(1-\frac{\alpha\overline{D}_{d_\beta}+\gamma\bar{D}}{\overline{D}_{d_\beta}-\bar{D}}\right) < (1-\alpha)P_0 < P_0$. □

A.3 Proof of Theorem 2.1

Proof. For given discount and premium parameters $\alpha \in [0,1]$ and $\gamma \geq 0$, we show that under conditions (a) and (b) in the statement of the Theorem there exists a **unique utility maximizing two-price contract**. In order to show this, we look for the two-price contract that satisfies property (i) and (ii) in Lemma 2.1. By Lemma 2.2, we can focus the search on the family of candidate contracts that satisfies risk reduction and extra-payment with equality, and show that there is a unique breakpoint that results in a contract (from the family) with demand break even point d_β. Thus, using the functional form of the prices given in Lemma 2.2, we first show that given condition (b), there exists a unique breakpoint that results in demand break even point d_β. This breakpoint is the unique solution to a specific equation. Then, we compute the optimal prices in closed form and show that they are uniquely determined.

Hereafter, we consider a two-price contract from the family of candidates. In order to induce demand break even point d_β, the breakpoint of the new contract

must satisfy

$$b^* = \frac{P_0 - P_2(b^*, \alpha, \gamma)}{P_1(b^*, \alpha, \gamma) - P_2(b^*, \alpha, \gamma)} d_\beta \tag{A.5}$$

Note that the order $P_2(b, \alpha, \gamma) \leq P_0 \leq P_1(b, \alpha, \gamma)$ implies that $b \leq d_\beta$. Then, using the functional form of the first price (equation (A.3)) given in the proof of Lemma 2.2, we can write

$$P_1(b, \alpha, \gamma) - P_2(b, \alpha, \gamma) = \frac{((1+\gamma)P_0 - P_2)\bar{D}}{\bar{D} - (\underline{D}_b - b)\underline{F}(b)} \tag{A.6}$$

We can plug (A.6) back into equation (A.5) to obtain

$$P_2(b, \alpha, \gamma) = P_0 \left(\frac{d_\beta(\bar{D} - (\overline{D}_b - b)\overline{F}(b)) - (1+\gamma)\bar{D}b}{d_\beta(\bar{D} - (\overline{D}_b - b)\overline{F}(b)) - \bar{D}b} \right) \tag{A.7}$$

We equal the latter expression to the second price functional form detailed in equation (A.4) for the case $b \leq d_\beta$. After rearranging terms, we obtain that the breakpoint of the utility maximizing contract with demand break even point d_β must satisfy

$$\begin{aligned} &\frac{(\bar{D} - (\overline{D}_b - b)\overline{F}(b))}{b} = \frac{\bar{D}}{d_\beta} + \frac{\gamma}{\alpha} \frac{\bar{D}(\overline{D}_{d_\beta} - d_\beta)}{\overline{D}_{d_\beta} d_\beta} \\ \Leftrightarrow &\int_0^b \left(1 - \frac{t}{b}\right) f(t)dt = 1 - \frac{\bar{D}}{d_\beta} - \frac{\gamma}{\alpha} \frac{\bar{D}(\overline{D}_{d_\beta} - d_\beta)}{\overline{D}_{d_\beta} d_\beta} \end{aligned} \tag{A.8}$$

This equation admits a non-negative solution if and only if the discount and premium parameters $\alpha \in [0, 1]$ and $\gamma \geq 0$ satisfy condition (b) in the statement of the Theorem. Moreover, this solution is unique. To see this, observe that the left hand side is continuous and increasing in b (its derivative is $\int_0^b \frac{t}{b^2} f(t)\, dt > 0$). In addition, the left hand side is zero at $b = 0$, and as $b \to d_\beta$, it approaches $\beta\left(1 - \frac{\underline{D}_{d_\beta}}{d_\beta}\right)$. The right hand side, on the other hand, is a non-negative constant (under condition (b)) which is smaller than $\left(1 - \frac{\bar{D}}{d_\beta}\right) < \beta\left(1 - \frac{\underline{D}_{d_\beta}}{d_\beta}\right)$. Hence, there is a unique solution for this equation in the interval $[0, d_\beta)$. Let us denote this breakpoint by $b^* = b^*(\alpha, \gamma)$.

To characterize the resulting prices, we can plug equation (A.8) into (A.7) to obtain the following expression for the second price $P_2(b^*, \alpha, \gamma) = \frac{P_0((1-\alpha)\overline{D}_{d_\beta} - d_\beta)}{\overline{D}_{d_\beta} - d_\beta}$. Then, replacing this back, together with (A.8), in the definition of the first price

(equation (A.3)), we obtain $P_1(b^*, \alpha, \gamma) = P_0\Big(1 + \frac{\alpha \overline{D}_{d_\beta}}{\overline{D}_{d_\beta} - d_\beta}\Big(\frac{d_\beta}{b^*} - 1\Big)\Big)$.

Observe that the second price is non-negative if condition (a) in the statement of the Theorem is satisfied. The uniqueness of the contract is guaranteed by the decreasing behavior of the prices in the interval $[0, d_\beta]$ stated in Lemma 2.2.

Finally, we conclude that, given the existence of a utility maximizing two-price contract, the SP does not gain anything by considering more complex contracts (i.e., contracts with multiple breakpoints). Hence, the SP can restrict to two-price incremental discount contracts in order to ensure the acceptable levels of risk quoted by the SR while obtaining his maximum possible utility. □

A.4 Proof of Corollary 2.1

Proof. Under the same assumptions as in Theorem 2.1, the SP will offer the two-price utility maximizing contract, as long as the discount and premium parameters are such that, the breakpoint and prices are non-negative, and the resulting SP's utility is at least his reservation utility.

The conditions on $\alpha \in [0, 1]$ and $\gamma \geq 0$ in Theorem 2.1 induce non-negative breakpoint and second price, that is

- $b^* \geq 0 \Leftrightarrow \gamma \frac{\bar{D}}{\overline{D}_{d_\beta}} \frac{(\overline{D}_{d_\beta} - d_\beta)}{(d_\beta - \bar{D})} \leq \alpha$
  Where the equivalence follows from equation (2.7)

- $P_2(b^*, \alpha, \gamma) \geq 0 \Leftrightarrow \alpha \leq 1 - \frac{d_\beta}{\overline{D}_{d_\beta}}$
  Where the equivalence follows from the second price closed from solution in Theorem 2.1.

Thus, the SP will offer a contract if the premium and discount are in the feasible region

$$J = \left\{ (\alpha, \gamma) | 0 \leq \gamma,\ \gamma \frac{\bar{D}}{\overline{D}_{d_\beta}} \frac{(\overline{D}_{d_\beta} - d_\beta)}{(d_\beta - \bar{D})} \leq \alpha \leq 1 - \frac{d_\beta}{\overline{D}_{d_\beta}} \right\}$$

Finally, the SP will participate in the agreement, if in addition to $(\alpha, \gamma) \in J$, the utility he obtains is at least his reservation utility, that is $EU^*_{SP}(\alpha, \gamma) \geq R_{SP}$. □

A.5 Proof of Theorem 2.2

Proof. The proof of the Theorem is organized in steps (i)-(iii). In step (i) we show that under conditions (a) and (b) of the Theorem, the SR's optimization problem is feasible, and express the feasibility region in terms of the discount and premium variables. In step (ii), we show that at optimality the SP utility participation constraint (2.10) is tight. This allow us to write the premium as a closed form expression of the discount. By restricting the SR problem to the set of feasible solutions that meet (2.10) with equality, we can rewrite the SR's problem as an alternative optimization problem that only depends on the discount variable, which is constrained to be in an interval. To conclude in step (iii), we show that the objective of the reformulated problem is strictly concave in the discount.

We assume that the SP is loss averse and his risk behavior is modeled by a piece-wise linear utility function. As the leader, the SR anticipates the outcome of the SP problem (lower level), thus, for $(\alpha, \gamma) \in J$, the SP's utility maximizing new contract corresponds to the one specified in equation (2.9).

(i) Assume that condition (a) and (b) are satisfied. The problem $\mathbf{P^{SR}}$ is feasible if we can find $(\alpha, \gamma) \in J$ that also satisfy the SP utility participation constraint (2.10). Mathematically,

$$EU^*_{SP}(\alpha, \gamma) \geq R_{SP} \;\Leftrightarrow\; \alpha(\nu_{SP} - 1)\frac{\overline{D}_{d_\beta}(1-\beta)}{\bar{D}} + \frac{R_{SP}}{P_0\bar{D}} \leq \gamma$$

This, together with the premium bound specified in J, restrict the premium variable to

$$\alpha(\nu_{SP} - 1)\frac{\overline{D}_{d_\beta}(1-\beta)}{\bar{D}} + \frac{R_{SP}}{P_0\bar{D}} \leq \gamma \leq \alpha\frac{\overline{D}_{d_\beta}}{\bar{D}}\frac{(d_\beta - \bar{D})}{(\overline{D}_{d_\beta} - d_\beta)} \tag{A.9}$$

In this last expression, the upper bound is greater or equal to the lower bound if condition (a) is true, and imply that the discount variable is above a specific threshold. Putting this threshold together with the discount upper bound (from

J) implies

$$\underbrace{\frac{R_{SP}}{P_0\left((d_\beta - \bar{D}) - (\nu_{SP}-1)(\overline{D}_{d_\beta} - d_\beta)(1-\beta)\right)}\left(1 - \frac{d_\beta}{\overline{D}_{d_\beta}}\right)}_{\alpha^{min}} \leq \alpha \leq \underbrace{1 - \frac{d_\beta}{\overline{D}_{d_\beta}}}_{\alpha^{max}} \tag{A.10}$$

The last set of bounds in the discount is only true under condition (a) and (b). Thus, under conditions (a) and (b) the inequalities (A.9) and (A.10) define the feasible region of problem $\mathbf{P^{SR}}$, which corresponds to a polyhedron. To simplify notation, hereafter we use $[\alpha^{min}, \alpha^{max}]$ as the feasible range for the discount variable.

(ii) We now show that at optimality the utility participation constraint (2.10) is tight.

Let us rewrite the objective function of problem $\mathbf{P}^{SR}$. To simplify the notation, we omit the dependence of the breakpoint and prices of the new contract in (α, γ), and denote them as (b^*, P_1^*, P_2^*). Given the optimal contract specified in Theorem 2.1, and that its induced demand break even point $d^* = d_\beta$, the utility of the SR can be written as

$$\begin{aligned}
\mathrm{E}[U_{SR}(C^{std}(D) - C^{new*}(D,\alpha,\gamma))] &= \int_0^{b^*} u_{SR}\Big(-\frac{\alpha P_0 \overline{D}_{d_\beta}}{\overline{D}_{d_\beta} - d_\beta}\Big(\frac{d_\beta}{b^*} - 1\Big)t\Big) f(t)\, dt \\
&+ \int_{b^*}^{d_\beta} u_{SR}\Big(-\frac{\alpha P_0 \overline{D}_{d_\beta}}{\overline{D}_{d_\beta} - d_\beta}(d_\beta - t)\Big) f(t)\, dt \\
&+ \int_{d_\beta}^{\infty} u_{SR}\Big(\frac{\alpha P_0 \overline{D}_{d_\beta}}{\overline{D}_{d_\beta} - d_\beta}(t - d_\beta)\Big) f(t)\, dt
\end{aligned} \tag{A.11}$$

The latter expression is derived by using the functional form of the optimal prices from equation (2.8) in Theorem 2.1 to write $P_0 - P_1^* = -\frac{\alpha P_0 \overline{D}_{d_\beta}}{\overline{D}_{d_\beta} - d_\beta}\Big(\frac{d_\beta}{b^*} - 1\Big)$ and $P_0 - P_2^* = \frac{\alpha P_0 \overline{D}_{d_\beta}}{\overline{D}_{d_\beta} - d_\beta}$.

From equation (A.11), note that the SR's expected utility does not depend on the premium variable directly, but through the breakpoint $b^*(\alpha, \gamma)$ (solution

to equation (2.7)). We also observe that the SR's expected utility is increasing in the breakpoint b^*. To see this, we recall that the SR's increasing loss aversion behavior is modeled by $u_{SR}(x) = -\nu_{SR}(-x)^{\theta_{SR}}$ with $\theta_{SR} > 1$ and $\nu_{SR} > 0$ for $x < 0$. Specifically, we have that $\frac{\partial \mathrm{E}[U_{SR}]}{\partial b^*} = \int_0^{b^*} u'_{SR}\Big(- \frac{\alpha P_0 \overline{D}_{d_\beta}}{\overline{D}_{d_\beta} - d_\beta}\Big(\frac{d_\beta}{b^*} - 1\Big)t\Big) \left(\frac{\alpha P_0 \overline{D}_{d_\beta}}{\overline{D}_{d_\beta} - d_\beta}\right) \frac{d_\beta}{b^{*2}} t f(t)\, dt > 0$.

In order to show that the utility participation constraint is tight at optimality, we first need to determine how $b^*(\alpha, \gamma)$ behaves as a function of the discount and premium variables. We recall that the optimal breakpoint satisfies equation (2.7). The left hand side of this equation is increasing in b^*, and the right hand side in increasing in α and decreasing in γ. Hence, the unique b^* that satisfies this equation is increasing in the discount α and decreasing in the premium γ. This, together with the previous observation that the expected utility of the SR is increasing in b^*, allow us to conclude that the SP's utility participation constraint must be tight at optimality.

In more detail, let us suppose that there exists an optimal solution $(\tilde{\alpha}, \tilde{\gamma})$, such that the utility participation constraint (2.10) is not tight. That is, the lower bound in (A.9) is not tight. Thus, by decreasing $\tilde{\gamma}$ by $\epsilon > 0$ small enough, such that $\tilde{\gamma} - \epsilon$ is still feasible, the new breakpoint is such that $b^*(\tilde{\alpha}, \tilde{\gamma} - \epsilon) > b^*(\tilde{\alpha}, \tilde{\gamma}) > 0$. Consequently, by decreasing the premium $\tilde{\gamma}$ slightly, the SR's expected utility is strictly increased. This contradicts the optimality of $(\tilde{\alpha}, \tilde{\gamma})$. Hence, any optimal solution of problem $\mathbf{P^{SR}}$ must satisfy the utility participation constraint with equality, namely,

$$\gamma^*(\alpha) = \alpha(\nu_{SP} - 1)\frac{\overline{D}_{d_\beta}(1-\beta)}{\bar{D}} + \frac{R_{SP}}{P_0 \bar{D}} \tag{A.12}$$

This last expression shows that the premium is uniquely determined by the discount.

(iii) Finally and given (ii), we show that the objective of problem $\mathbf{P^{SR}}$ in concave, hence, there is a unique optimal solution.

By (ii), we replace $\gamma^*(\alpha)$, and notice that the breakpoint b^* becomes independent of the premium variable. Thus, problem $\mathbf{P^{SR}}$ becomes a single variable (discount α) optimization problem and the feasible region corresponds to (A.10). We show that the objective of this single variable optimization problem is concave in the discount variable α. The SR's optimization problem is reduced to the following formulation

$$\begin{aligned}
&\max_{\alpha,b^*} \mathrm{E}[U_{SR}(C^{std}(D) - C^{new*}(D,\alpha,\gamma^*(\alpha)))] \\
&s.t. \quad \int_0^{b^*} \left(1 - \frac{t}{b^*}\right) f(t)dt = \\
&\qquad 1 - \frac{\bar{D}}{d_\beta} - (\nu_{SP} - 1)\frac{(\overline{D}_{d_\beta} - d_\beta)(1-\beta)}{d_\beta} - \frac{1}{\alpha}\frac{(\overline{D}_{d_\beta} - d_\beta)R_{SP}}{P_0\overline{D}_{d_\beta}d_\beta} \qquad \text{(A.13)} \\
&\qquad \alpha \in [\alpha^{min}, \alpha^{max}]
\end{aligned}$$

Effectively, this optimization problem is a single variable optimization problem. The equality constraint (A.13) corresponds to equation (2.7) after replacing for $\gamma^*(\alpha)$. The breakpoint $b^* \equiv b(\alpha)$ is a function of the discount variable and is uniquely determined by the equality constraint (A.13). We use this constraint to determine the first order behavior of the breakpoint as a function of the discount level α by taking total derivatives in both sides, $b'(\alpha) = \frac{b(\alpha)^2}{\alpha^2}\frac{R_{SP}}{d_\beta H \int_0^{b(\alpha)} tf(t)\,dt} > 0$, where $H = \frac{P_0\overline{D}_{d_\beta}}{\overline{D}_{d_\beta} - d_\beta}$ is a constant. The objective of the SR's optimization problem can be written as $\mathrm{E}[U_{SR}(C^{std}(D) - C^{new*}(D,\alpha,\gamma^*(\alpha)))] = -\nu_{SR}(\alpha H)^\theta(I_1^\theta(\alpha) + I_2^\theta(\alpha)) + \alpha H(\overline{D}_{d_\beta} - d_\beta)(1-\beta)$. In this last expression, we consider the SR's utility function previously specified. To simplify notation, we omitted the lower script SR from the loss aversion parameter θ, and also introduce the notation $I_1^\theta(\alpha) = \int_0^{b(\alpha)} \left(\frac{d_\beta}{b(\alpha)} - 1\right)^\theta t^\theta f(t)\,dt$ and $I_2^\theta(\alpha) = \int_{b(\alpha)}^{d_\beta}(d_\beta - t)^\theta f(t)\,dt$.

To show that the objective function is concave in the discount variable α, we

take second order derivatives.

$$\begin{aligned}
&\mathrm{E}[U_{SR}(C^{std}(D) - C^{new*}(D,\alpha,\gamma^*(\alpha)))]'' = \\
&-\nu_{SR}H^\theta\alpha^{(\theta-2)}\left(\theta(\theta-1)(I_1^\theta(\alpha)+I_2^\theta(\alpha))\right. \\
&\left.+2\theta\alpha(I_1^\theta(\alpha)+I_2^\theta(\alpha))' + \alpha^2(I_1^\theta(\alpha)+I_2^\theta(\alpha))''\right)
\end{aligned} \tag{A.14}$$

Using Leibneiz's rule, and the definition of $b'(\alpha)$, we obtain $(I_1^\theta(\alpha)+I_2^\theta(\alpha))' = -\theta\frac{R_{SP}}{\alpha^2 H}\frac{I_1^\theta(\alpha)}{I_1^1(\alpha)}$, and
$(I_1^\theta(\alpha)+I_2^\theta(\alpha))'' = \theta\frac{R_{SP}}{\alpha^3 H I_1^1(\alpha)^2}\left(2I_1^\theta(\alpha)I_1^1(\alpha) + (\theta-1)\frac{R_{SP}}{\alpha H}I_1^\theta(\alpha)\right.$
$\left.+\alpha(d_\beta - b(\alpha))f(b(\alpha))b'(\alpha)\Big(I_1^\theta(\alpha) - (d_\beta - b(\alpha))^{(\theta-1)}I_1^1(\alpha)\Big)\right)$.

Plugging these back into the second derivative of the objective (A.14),
$\mathrm{E}[U_{SR}(C^{std}(D) - C^{new*}(D,\alpha,\gamma^*(\alpha)))]'' =$

$$\begin{aligned}
&-\nu_{SR}H^\theta\theta\alpha^{(\theta-3)}\left(\underbrace{(\theta-1)I_1^\theta(\alpha) - 2(\theta-1)\frac{R_{SP}}{H}\frac{I_1^\theta(\alpha)}{I_1^1(\alpha)} + (\theta-1)\frac{R^{SP^2}}{\alpha H^2}\frac{I_1^\theta(\alpha)}{I_1^1(\alpha)^2}}_{Q}\right. \\
&+\underbrace{(\theta-1)I_2^\theta(\alpha)}_{>0} \\
&\left.+\frac{R_{SP}}{HI_1^1(\alpha)^2}\alpha(d_\beta - b(\alpha))f(b(\alpha))b'(\alpha)\underbrace{\Big(I_1^\theta(\alpha) - (d_\beta - b(\alpha))^{(\theta-1)}I_1^1(\alpha)\Big)}_{\geq 0}\right)
\end{aligned}$$

In order to conclude that $\mathrm{E}[U_{SR}(C^{std}(D) - C^{new*}(D,\alpha,\gamma^*(\alpha)))]'' < 0$, we notice that $Q = \frac{(\theta-1)I_1^\theta(\alpha)}{\alpha H^2 I_1^1(\alpha)^2}\left(\alpha H I_1^1(\alpha) - R_{SP}\right)^2 \geq 0$. Hence, given the increasing loss aversion assumption, i.e., $\theta > 1$, the objective function is a strictly concave function of the discount variable. Thus, since the discount feasible region is an interval, the SR's optimization problem admits a unique optimal solution in this interval.

Given the concavity of the objective, we obtain the unconstrained utility max-

imizing discount, $\tilde{\alpha}$ by the first order condition

$$\mathrm{E}[U_{SR}(C^{std}(D) - C^{new*}(D, \alpha, \gamma^*(\alpha)))]'|_{\alpha=\tilde{\alpha}} = 0$$

$$\Leftrightarrow \tilde{\alpha} = \frac{1}{H}\left(\frac{(\overline{D}_\beta - d_\beta)(1-\beta)}{\nu_{SR}\theta_{SR}\left(I_1^{\theta_{SR}}(\tilde{\alpha}) + I_2^{\theta_{SR}}(\tilde{\alpha}) - \frac{R_{SP}}{\tilde{\alpha}H}\frac{I_1^{\theta_{SR}}(\tilde{\alpha})}{I_1^1(\alpha)}\right)}\right)^{1/(\theta_{SR}-1)}$$

Thus, as for any concave function optimized over an interval, the optimal solution corresponds to

$$\alpha^* = \begin{cases} \tilde{\alpha} & \tilde{\alpha} \in [\alpha^{min}, \alpha^{max}] \\ \alpha^{max} & \tilde{\alpha} > \alpha^{max} \\ \alpha^{min} & \tilde{\alpha} < \alpha^{min} \end{cases}$$

□

A.6 Proof of Corollary 2.2

Proof. Proof:

In the case $R_{SP} = 0$, the unconstrained discount reduces to

$$\tilde{\alpha} = \frac{\overline{D}_{d_\beta} - d_\beta}{P_0\overline{D}_{d_\beta}}\left(\frac{(\overline{D}_\beta - d_\beta)(1-\beta)}{\nu_{SR}\theta_{SR}\left(\int_0^{b^*}((\frac{d_\beta}{b^*} - 1)t)^{\theta_{SR}}f(t)\,dt + \int_{b^*}^{d_\beta}(d_\beta - t)^{\theta_{SR}}f(t)\,dt\right)}\right)^{\frac{1}{(\theta_{SR}-1)}}$$

Where b^* is unique solution to (A.13) after replacing $R_{SP} = 0$. Note that the unconstrained discount might not be feasible. In order for it be feasible, it must satisfy (A.10), otherwise the equilibrium discount corresponds to one of the extremes values of the discount feasible range.

□

A.7 Proof of Theorem 2.3

Proof. Proof: We begin by characterizing the distribution of the normalized payments under the new and standard contracts. Based on this, we show that these distributions cross in exactly one point (in the common support), and using the fact that they have the same mean, we conclude that they satisfy the Dilation and Lorenz stochastic orders.

1. Dilation order

   Based on Definition 2.3, consider the normalized random variables $C^{std}(D) - \bar{C}^{std}$ and $C^{new}(D) - \bar{C}^{new}$ (both with finite mean). We derive the distribution of the normalized random variables based on the demand CDF (F)

$$F^{std}(t) = P(C^{std}(D) - \bar{C}^{std} \leq t) = F\left(\frac{t - P_0\bar{D}}{P_0}\right), \; P_0\bar{D} \leq t$$

$$F^{new}(t) = P(C^{new}(D) - \bar{C}^{new} \leq t)$$

$$= \begin{cases} F\left(\frac{t-\bar{C}^{new}}{P_1}\right) & \bar{C}^{new} \leq t \leq P_1 b + \bar{C}^{new} \\ F\left(\frac{t-\bar{C}^{new}}{P_2} - \frac{(P_1-P_2)}{P_2}b\right) & P_1 b + \bar{C}^{new} < t \end{cases}$$

   We proceed to show that at equilibrium the CDFs cross in exactly one point, and at this point, the CDF of the new contract it does it from below. Recall our initial assumption on the demand CDF being monotonically increasing. In addition, recall that at equilibrium the expected payment under the new contract is larger than the standard contract, i.e., $P_0\bar{D} < \bar{C}^{new}$, by equation (2.11). Thus, for $P_0\bar{D} \leq t < \bar{C}^{new}$, we have that $0 = F^{new}(t) < F^{std}(t)$. Similarly, for $\bar{C}^{new} \leq t \leq P_1 b + \bar{C}^{new}$, we also have that $F^{new}(t) < F^{std}(t)$ since $P_0 < P_1$. Finally, since both normalized random variables have zero mean, we know that the CDFs must cross at some point $\tilde{t} > P_1 b + \bar{C}^{new}$. The CDFs intersect at $\tilde{t} = \frac{P_0}{(P_1-P_2)}\left((P_1 - P_2)b + \bar{C}^{new} - P_2\bar{D}\right)$.

   It is straightforward to see that $\tilde{t} > P_1 b + \bar{C}^{new}$. Then, there is a unique crossing point, and at this, the new contract CDF cross the CDF of the standard contract

from below. In order to show the convex order between the normalized payment random variables, we observe that they satisfied condition 3.A.8 in [78].

Let $0 \leq x \leq \tilde{t}$, we have that $\int_0^x F^{new}(t)dt \leq \int_0^x F^{std}(t)dt$.

Let $x > \tilde{t}$. Then using the fact that the normalized distributions have the same mean,

$$\begin{aligned} & \int_0^\infty F^{new}(t)dt = \int_0^\infty F^{std}(t)dt \\ \Leftrightarrow \int_0^x F^{new}(t)dt + \int_x^\infty F^{new}(t)dt &= \int_0^x F^{std}(t)dt + \int_x^\infty F^{std}(t)dt \\ \Rightarrow \int_0^x F^{new}(t)dt &\leq \int_0^x F^{std}(t)dt \end{aligned}$$

The last inequality follows from $F^{new}(t) > F^{std}(t)\ \forall t > \tilde{t}$. Then, by Theorem 3.A.1 (b) in [78], the normalized payment are in convex order, or equivalently $C^{new} \leq_{dil} C^{std}$.

2. Lorenz order

Based on Definition 2.3, consider the normalized non-negative random variables $\frac{C^{std}(D)}{\bar{C}^{std}}$ and $\frac{C^{new}(D)}{\bar{C}^{new}}$, and their CDFs are given by

$$F^{std}(t) = P\left(\frac{C^{std}(D)}{\bar{C}^{std}} \leq t\right) = F\left(t\bar{D}\right),\ 0 \leq t$$

$$F^{new}(t) = P\left(\frac{C^{new}(D)}{\bar{C}^{new}} \leq t\right) = \begin{cases} F\left(\frac{t\bar{C}^{new}}{P_1}\right) & 0 \leq t \leq \frac{P_1 b}{\bar{C}^{new}} \\ F\left(\frac{t\bar{C}^{new}}{P_2} - \frac{(P_1-P_2)}{P_2}b\right) & \frac{P_1 b}{\bar{C}^{new}} < t \end{cases}$$

Now, we show that at equilibrium the new contract CDF cross the CDF of the standard contract in one point, and at this point, it does it from below. Recall our initial assumption on the CDF of the demand being monotonically increasing. Thus, for $0 \leq t < \frac{P_1 b}{\bar{C}^{new}}$, we have that $0 = F^{new}(t) < F^{std}(t)$. This is derived from the fact that $\bar{C}^{new} < P_1\bar{D}$ which is directly implied from $P_2 < P_0 < P_1$. Hence, since both normalized random variables have mean equal to one, we know that the CDFs must cross at some point $\tilde{t} > \frac{P_1 b}{\bar{C}^{new}}$. The

CDFs intersect at $\tilde{t} = \frac{(P_1 - P_2)b}{C^{new} - P_2 D}$. It is straightforward to see that $\tilde{t} > \frac{P_1 b}{C^{new}}$. Then, there is a unique crossing point, and at this, the new contract CDF cross the CDF of the standard contract from below. Following the same idea we did for the Dilation order, we can show that the normalized payment random variables satisfy condition 3.A.8 in [78]. Then, by Theorem 3.A.1 (b) in [78], the normalized payments are in convex order, or equivalently $C^{new} \leq_{Lorenz} C^{std}$.

□

Appendix B

Additional Material Chapter 3

B.1 Data description and estimation of parameters

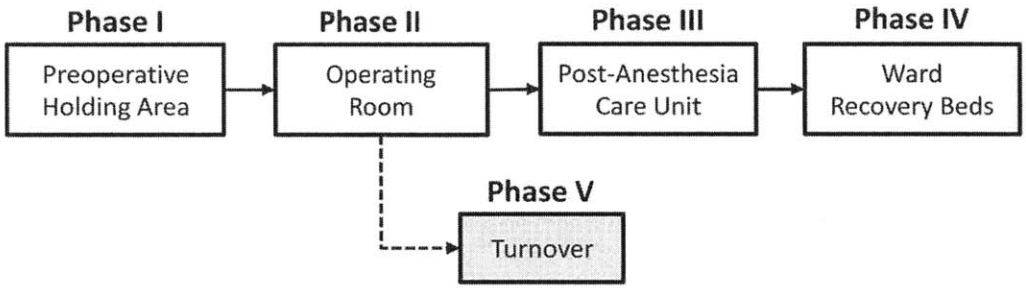
We summarize the methodology we used to estimate the parameters of the model in the application of our framework. The steps listed below do not need to be followed strictly, but they can be used as guidelines for different applications of our framework. In terms of data, we used historical financial, surgical records, and capacity data from the AMC and community hospitals (where available) from years 2009-2013. However, most of the parameters were estimated using the last complete year of data, that is, 2012 as a way to represent the current operations.

1. **Selection of procedure types**: Procedures types are identified by the AMC's internal surgical coding system, and we used *minimum volume* as the criteria to select procedures types. Specifically, we focus on surgical procedures across three surgical departments (general, colorectal,), and select those that were performed at least once a month according to the AMC's surgical volumes in 2012. Although this threshold is arbitrary, it allows us to have minimum data to compute statistics on the usage of resources, and revenue and cost. Infrequent procedure types tend to have highly variable resource consumption patterns, moreover, the benefit of including these procedures from a managerial perspective is unclear since they are rarely performed. Hence, we exclude them

to avoid introducing noise and skewing results. This criteria resulted in the selection of 57 surgical procedures types.

2. **Selection and classification of resources**: In order to identify the relevant resources for the studied procedures, we mapped the resources used along the surgical path. Figure B-1 shows the phases of the surgical path included in this study. Patients move along phases I-IV, but phase V is required for every patient after phase II is performed.

Figure B-1: Phases of surgical path.



Generally, resources can be of three types: equipment or supplies (type A), physical infrastructure (type B), and staff (type C). The choice of type A resources was based on *minimum usage* according to the AMC's surgical records. Specifically, items that are not critical, and used in less than 5% of the cases of each procedure type are excluded. Furthermore, we reduced this list by eliminating low-cost items that are included within the physical infrastructure (e.g., surgical table is part of the operating room). This selection process resulted in a final list of 197 type A resources. Type B resources correspond to operating rooms, preoperative and post-anesthesia bays, and ward beds. Type C resources broadly include different nurses, anesthesiologists, and surgeons. We assumed that most type C resources (except for the AMC's surgeons) are included as part of the physical infrastructure. The reason behind this is that, as it is a common practice in many hospitals, anesthesiologists and nurses are staffed on an aggregate per operating room basis. Moreover, preoperative and post-anesthesia bays also follow a similar staffing model. Thus, having one op-

erating room available in our application means that the surgical team (except for the surgeon) is guaranteed for that room. Hence, the only type C resource that is explicitly included in the model is the surgeon.

Within the surgeon resource, there are different sub-types depending on surgeons' skills. We designed surgeons' classes, such that each individual surgeon belongs to exactly one class. Each class corresponds to a set of surgical sub-specialties. We start by identifying individual surgeons who can perform the studied procedures. In addition, we group procedure types into sub-specialties, such that each procedure type belongs to exactly one sub-specialty; individual surgeons can perform procedures from multiple sub-specialties. We classify individual surgeons into classes based on the sub-specialties they can perform. For example, John and Jane specialize in procedures from two sub-specialties *bariatric* and *colorectal*. Judy, on the other hand, masters procedures in *colorectal* and *endocrine* sub-specialties. We define two classes, {bariatric and colorectal} and {colorectal and endocrine}. John and Jane belong to the first class, and Judy to the second. Notice that colorectal procedures can be provided by surgeons from both classes. We introduced this modeling in order to cluster surgeons skills which will ultimately facilitate the estimation of available operating time for the different procedure types. In our study we have a total of 20 surgeons, 7 sub-specialties, and 10 surgeon classes.

In addition, we also identified resources according to the general model classification, that is, whether they are *fixed*, *flexible*, and *substitutable*. Thus, resource types A and B are assumed fixed to a specific location, while the surgeons (type C) is flexible. This means that surgeons' time could be allocated to different hospitals across the network. In addition, we consider surgeons (different classes) as a substitutable resource since a single procedure can be performed by surgeons from different classes.

3. **Estimation of resource usage**: Resource consumption is defined in *time equivalent units* (i.e., the quantity of resource used times the length of the time

that the resource is used). We assume that resources required in a specific phase, are used for the entire duration of the phase.

Phase duration is defined as the total time that the patient spent on the specific phase. For every procedure type, we computed the empirical distribution of the duration of each phase according to the AMC 2012 volumes. For phases I-III, we observed that the durations for different procedures differ in variability, and right skew. Thus, we used the median of the empirical duration as the typical duration. The duration of phase IV is modeled as the typical length of stay (LOS). Since some procedures are performed in both inpatients and outpatients settings, using median duration does not represent the typical duration of either subset of patients. Hence, for each procedure type we estimated the empirical LOS as the weighted average of the two subset of patients based on the AMC 2012 volumes,

$$\text{Typical LOS} = \frac{\text{Inpatient cases}}{\text{Total cases}} \times \text{Inpatient median LOS}$$

We assumed that the duration of phases I-IV is the same across locations. Conversely, the duration of the turnover (phase V) is not procedure specific but location dependent, and it is defined according to the scheduling standards at each individual hospital (e.g., 30 minutes at the AMC, 45 minutes at the community).

Resource quantity corresponds to amount of resource used in each phase. The typical quantity is modeled by the average quantity used according to the AMC 2012 volumes.

Although our general model allows for different resource usage at each location, we assumed that the typical usage is the same across hospitals. According to the local nursing teams, required resources and usage are inherent to the procedure and are not expected to change significantly based on location. Differences in resource requirements were exceptions and we handled them individually for the specific procedures and locations. Thus, the typical resource usage (u_{parl} in

general formulation) is defined as follow, for a given phase a, the typical usage of resource r by procedure p is

$$\begin{aligned}\text{Typical usage}_{par} = \text{Average resource } r \text{ used by procedure } p \text{ in phase } a \\ \times \text{Duration of phase } a\end{aligned}$$

Notice that phases can be interpreted as activities in the general formulation.

4. **Estimation of capacity**: At the AMC, the capacity and inventory of resources is readily acquired from the IT system. At the community hospitals, on the other hand, such a system does not exist, and we had to manually collect data with the assistance of the local nursing teams. Limiting the analysis to a subset of procedures types introduces a unique challenge in capacity modeling. Specifically, resources are used by all procedures types, including those that are not included in our model. To reconcile this discrepancy, we approximately segmented the capacity into the portion that is available for the studied procedures.

   We began by estimating the capacity of type B resources. The common practice in most academic hospitals is to split operating room time into segments, or blocks, that are assigned to individual surgeons. At the AMC, we use the 2012 block time allocation to determine the percentage of block time allocated to surgeons who can perform the studied procedures. This allocation is scaled further to account for the portion of time that these surgeons have historically spent operating cases of the studied procedures.

$$\begin{aligned}\text{Available OR capacity} = \text{Total capacity} \times \text{Surgeons block time share(\%)} \\ \times \text{Studied procedures share(\%)}\end{aligned}$$

   We used the same scaling to derive an approximate estimate of the available preoperative and recovery bays capacity. In the case of ward-beds capacity, we directly estimated the capacity utilized according to the AMC 2012 volumes,

and used this as the minimum capacity available and perform sensitivity analysis with respect to it. For the community hospitals, we directly asked the operating room managers to estimate the spare capacity for all the physical infrastructure ((B) resources).

The capacity of type A resources is more difficult to scale. Cleaning, transportation and set-ups consume usable time. Moreover, scaling based on volume or operating room utilization might not represent usage among different procedures. Therefore, we adjusted total capacity proportionally to block time share. Even though this is a conservative assessment, it is more realistic than assuming that the entire capacity is available for the studied procedures. For type A and B resources, we scaled capacity so this is measured in time equivalent units. The daily availability corresponds to the normal operating hours of the respective hospital. Preoperative bays and operating rooms are usually available between 9-10 hours, while recovery bays are usually available for longer hours (around 2 hours more than operating rooms). Ward-beds are available 24/7. For other resources, we assumed the same daily availability as for the operating room.

The surgeons operating time (resource type C) capacity is computed on a surgeon class basis based on the AMC 2012 records. Thus, if multiple surgeons belong to a class their capacity is pooled. To estimate individual surgeons operating time capacity, we began by computing the total operating time (TOT) spent on the studied procedures. TOT includes all the operating time, regardless of whether the case was performed within surgeons' own block time. This consideration is particularly relevant for those surgeons who do not own block time and only schedule cases through the waiting list. (Note that this is a conservative estimate of the time each surgeon could spend operating cases of the studied procedures.) In addition, we estimated individual surgeon's adjusted block time (ABT). This corresponds to the surgeon's individual block time allocation, but scaled by the share of surgeon's actual operating time spent on the studied procedures, plus the time from group blocks. Group blocks are

standard blocks that are shared by a group of surgeons, and we allocate this time based on historical usage. Again, ABT is a conservative estimate of the time surgeons could spent operating cases of the studied procedures. Finally, surgeon's available time is defined as the maximum between the actual time spent on studied procedures, and the adjusted block time. Then, the class time is just the pooled capacity across the surgeons in the class,

$$\text{Surgeon Class Time} = \sum_{i \in Surgeon\ Class} \max\{TOT_i, ABT_i\}$$

5. **Estimation of revenue and cost**: As we have done for other non-deterministic inputs, we consider typical profitability as the average (revenue - variable cost) collected by procedure type and location. We used payments as a proxy for revenue. Unfortunately, hospital payments are only recorded by patient encounter using a different coding system, ICD-9 procedure codes, that does not uniquely map to the internal coding system employed in our model. In order to compute estimates of revenue by procedure type, we created a mapping based on the empirical occurrence (weights) of ICD-9 codes using the AMC's payments data. Each encounter can potentially correspond to multiple procedure types (e.g., patients having multiple surgeries) and multiple ICD-9 codes, however, we can identify a primary ICD-9 code for each encounter. Consider a procedure type x, and an ICD-9 primary code y. We estimated the conditional probability of having a procedure of type x given that we observed the ICD-9 primary code y as

$$P(x|y) = \frac{P(x,y)}{P(y)} \sim \frac{\text{No. encounters of } x \text{ and } y}{\text{No. encounters of } y}$$

Thus, the revenue of a specific procedure type is simply the weighted average of the payments received by primary ICD-9 codes.

$$\text{Payment}(x) = \sum_{y \in \text{ICD-9}} P(x|y) \times \text{Avg. Payment}(y)$$

In the case of community hospitals, payments data is also indexed by ICD-9 codes, and we use the above mapping to estimate the payment by procedure type. Unfortunately, some ICD-9 are rarely performed in the community, or not at all, resulting in sample size issues. In order to have meaningful estimates of the average payments by ICD-9 in the community, we adjusted the average payment of ICD-9 codes that had less than 10 records using the AMC's payments. Thus, for an ICD-9 with sample size $n < 10$

$$\begin{aligned}\text{Adjusted Avg. COM payment} = \text{Avg. COM payment} \left(\frac{n}{10}\right) \\ +\gamma \text{ Avg. AMC payment} \left(\frac{10-n}{10}\right)\end{aligned}$$

Where $\gamma > 0$ is the average payment differential between the community, and the AMC, across ICD-9 codes with 10 or more records. For example, $\gamma = 0.8$ means that the payments in the community are on average 20% lower than at the AMC.

Cost data is also indexed by encounter and ICD-9 codes, but further broken down into specific cost departments. As previously described, we only include variable cost in our model. This cost accounts for surgical supplies and disposable equipment, pharmacy, and other minor expenses. Network capacity costs (fixed cost) are excluded from the model since they are basically committed well in advance and do not significantly vary with small changes in the case-mix. Instead, we model the consumption of this capacity explicitly in the model. We estimated average cost by procedure type based on the AMC 2012 cost records and we assume the cost in the community is the same.

6. **Estimation of demand**: We included two sources of demand; existing demand and leaked demand. Existing demand corresponds to the AMC 2012 volume for the set of studied procedures. We considered this as the baseline demand. Leaked demand was estimated by analyzing claims data of in-network patients that received care outside the network in 2012. The leaked demand

records are indexed by the Diagnosis-Related Groups (DRGs) coding systems; the HCFA/CMS-DRG for Medicare cases and the AP-DRG for non-Medicare cases. The mapping between this coding system and the hospital internal coding system is non-trivial because multiple DRGs can be assigned to a single procedure type. Similarly to the ICD-9 mapping, we created a mapping based on the empirical frequency of DRGs on existing demand. Consider a procedure type x and an DRG code z, then using existing volumes, we estimated the probability of having a procedure of type x given that we observed DRG z as

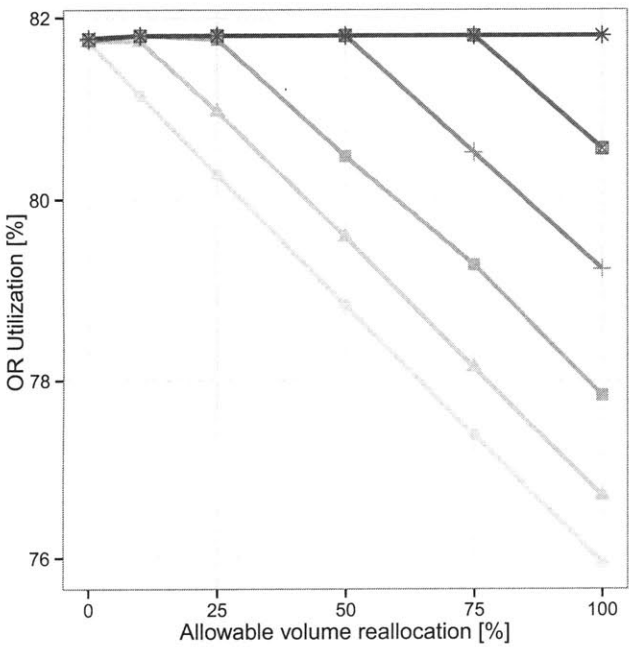
$$P(x|z) = \frac{P(x,z)}{P(z)} \sim \frac{\text{No. cases of } x \text{ and } z}{\text{No. cases of } z}$$

Thus, the unmet demand for a specific procedure type x is simply the weighted average of the leaked demand volumes

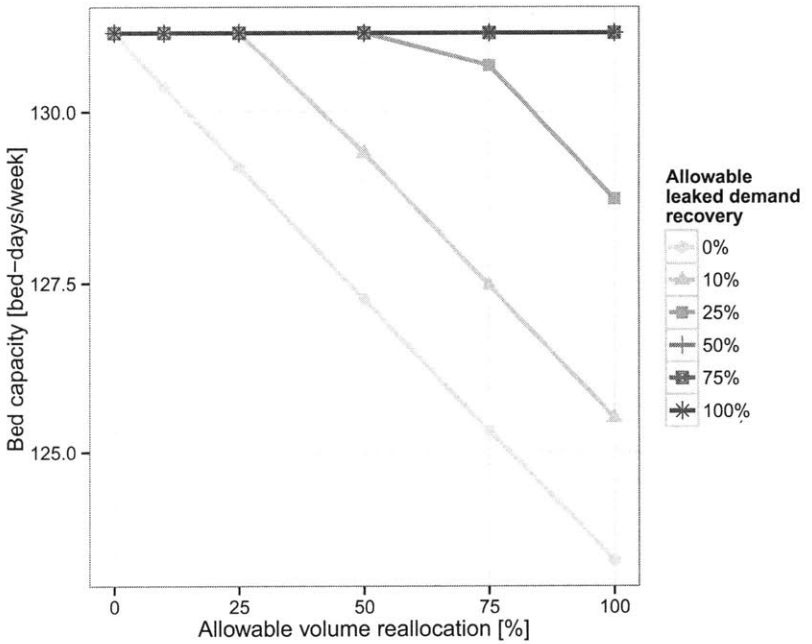
$$\text{Unmet demand}(x) = \sum_{z \in DRG} P(x|z) \times \text{Leaked demand}(z)$$

The estimates of existing, and leaked demand, are used to determine the various demand bounds in the model. Minimum network demand is defined as the existing volumes. Notice that in the past, this volume was entirely seen at the AMC, but in this study, we allowed for potential reallocation across the network, if profitable to do so. Maximum network demand corresponds to the existing demand plus some fraction of the leaked demand, and we do sensitivity analysis with respect to it. In addition, we used a minimum volume requirement at the AMC to control for how much of the existing demand could be reallocated to other hospitals in the network. For instance, if the minimum volume is exactly the existing demand, then no reallocation will be allowed in the optimal solution. To model different reallocation levels, we considered different fractions of the existing demand as the minimum volume requirement. In the case of community hospitals, no minimum volume is required and cases will be allocated to these hospitals only if larger benefit is obtained by doing so.

Figure B-2: AMC operating room utilization, and required ward-beds capacity.

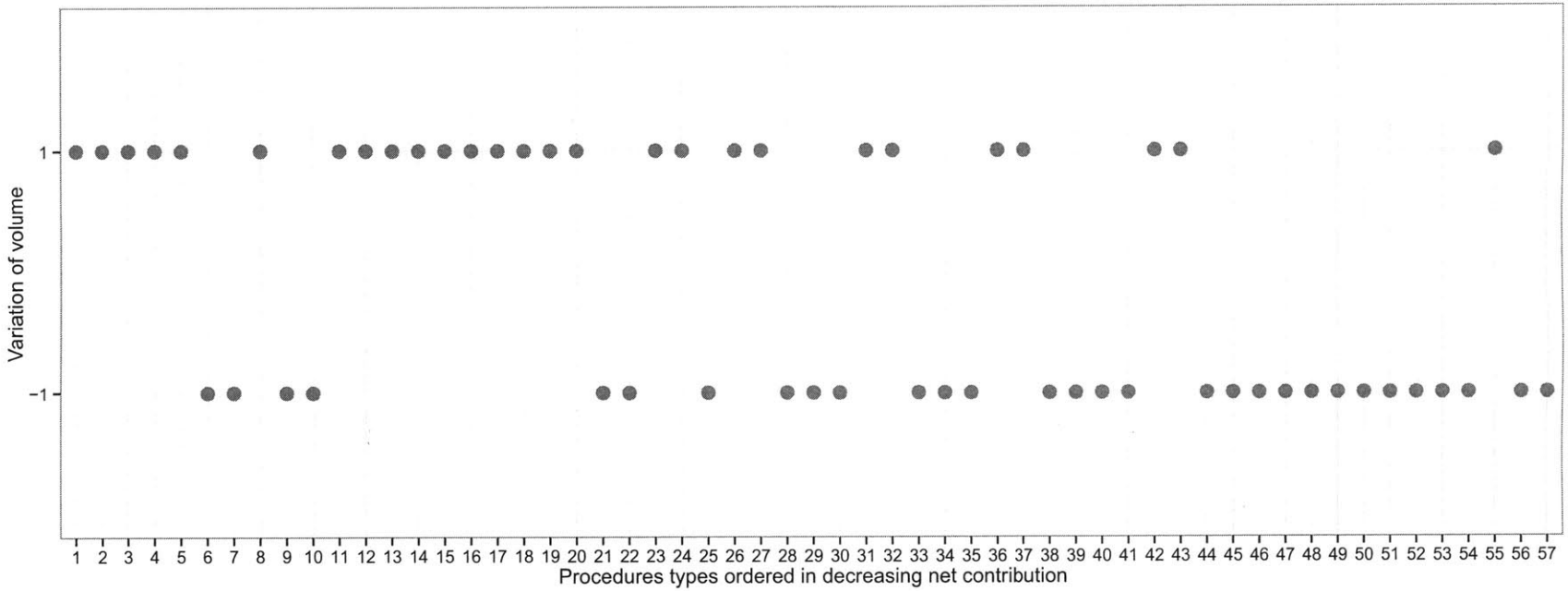


(a) Operating room utilization.



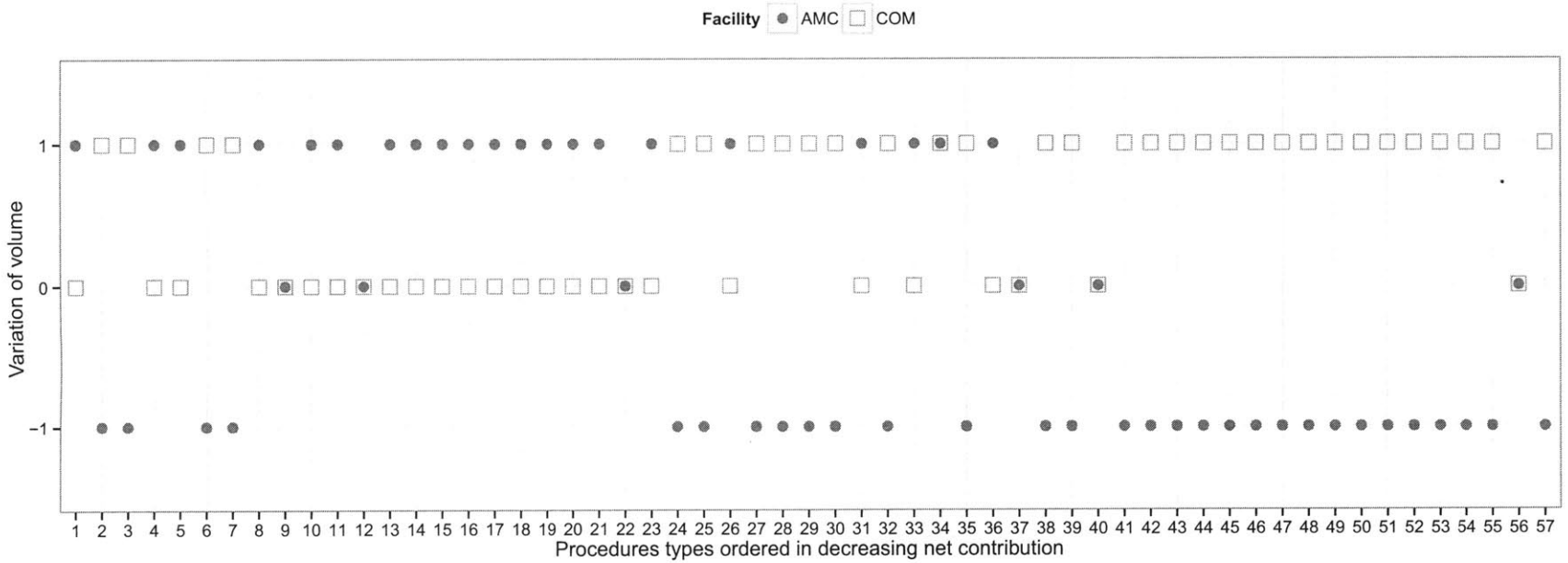
(b) Ward-beds capacity requirements.

Figure B-3: Changes in AMC's portfolio of services.



Note: Volume increase is represented by +1, and the decrease by -1. We assume that individual procedure types volume can vary in up to ±5%, and total volume stays within ±1% of the total baseline volume

Figure B-4: Changes in network's portfolio of services when recovering leaked demand.

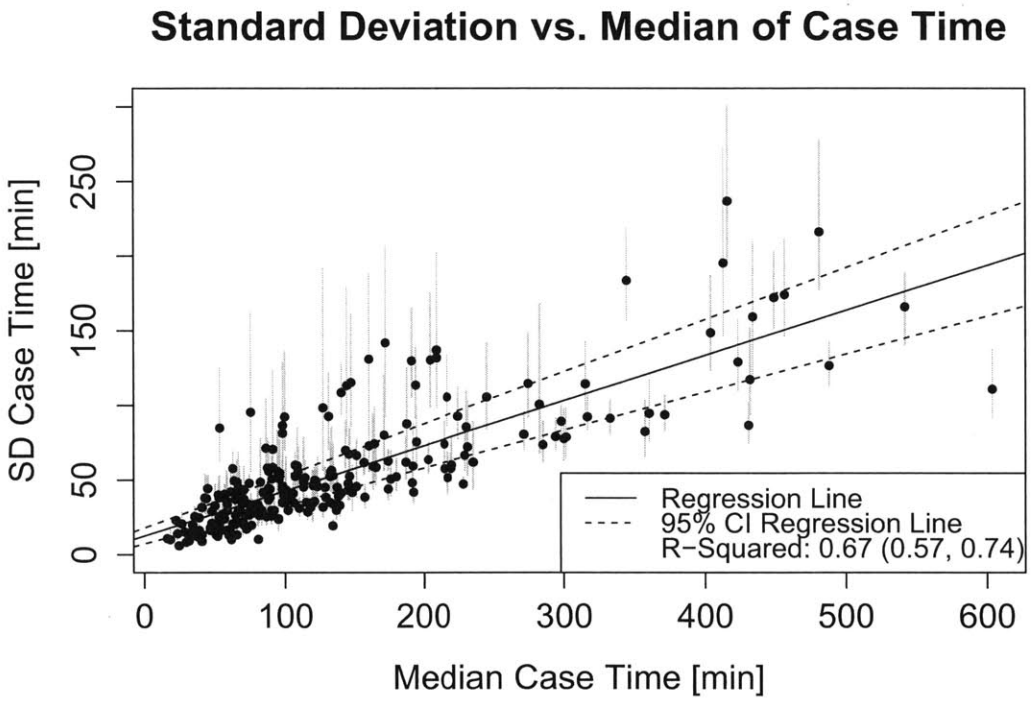


Note: Assume 10% reallocation, and full leaked demand recovery. Volume changes are measured relative to the baseline volumes. Positive volume changes are coded as +1, negative as -1, and no change as 0. AMC is represented by filled circles and the community as unfilled squares. When AMC and community have no change in volume (both at zero), it means that there is no leaked demand for the specific procedure.

Appendix C

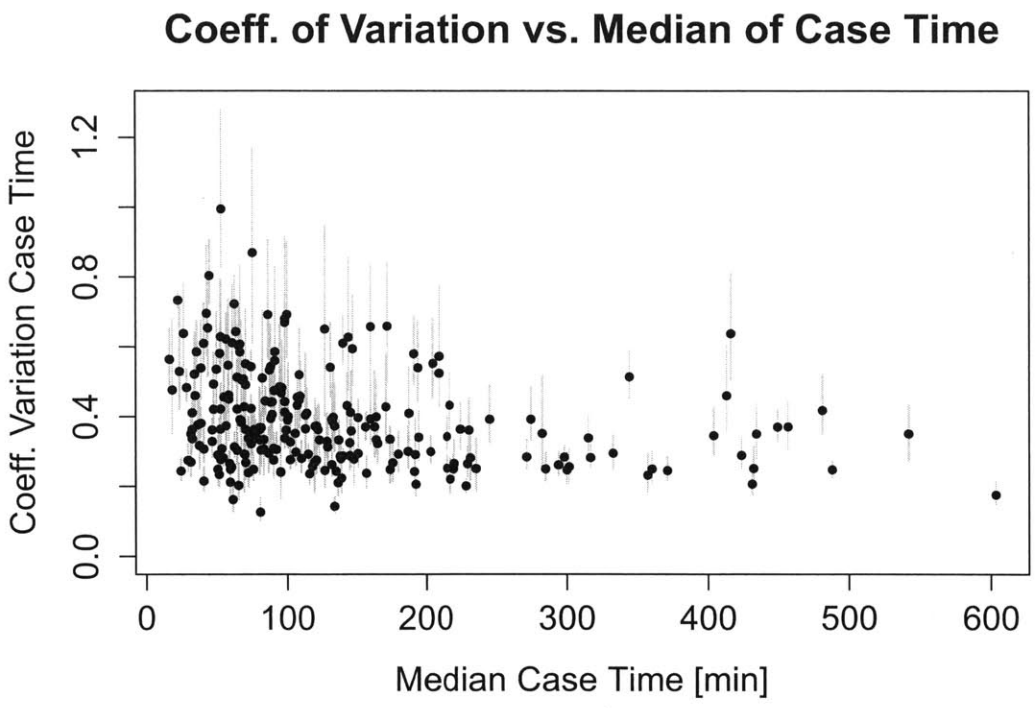
Figures and Tables Chapter 4

Figure C-1: Standard deviation vs. median case time.



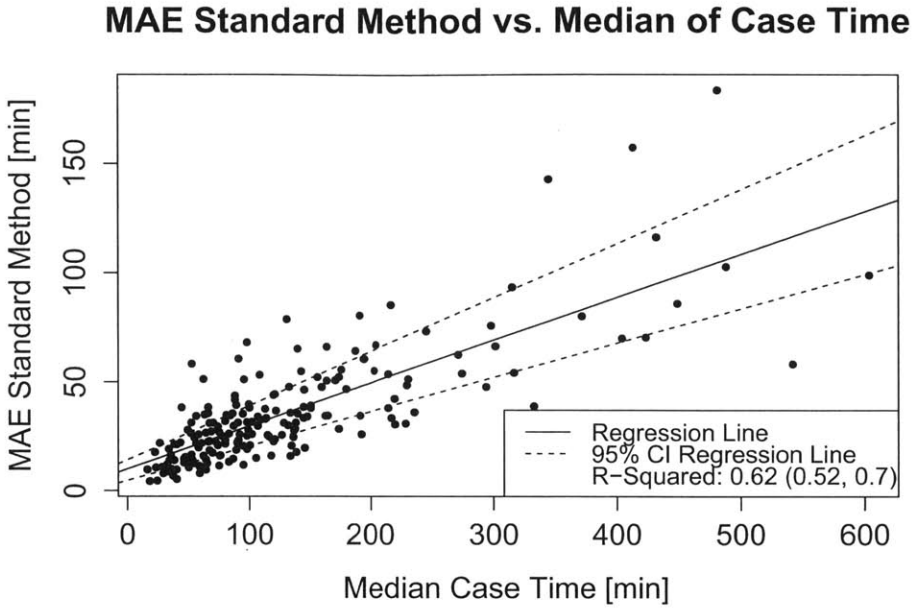
Note: Linear regression slope is 0.30 (0.26, 0.35), intercept 12.56 (7.77, 17.61). The 95% CI linear regression and R-Squared were computed via bootstrapping (BCa, 1000 replications). Each point corresponds to a procedure type, and the vertical bars to the 95% bootstrap CI of the standard deviation and coefficient of variation, respectively.

Figure C-2: Coefficient of variation vs. median case time.

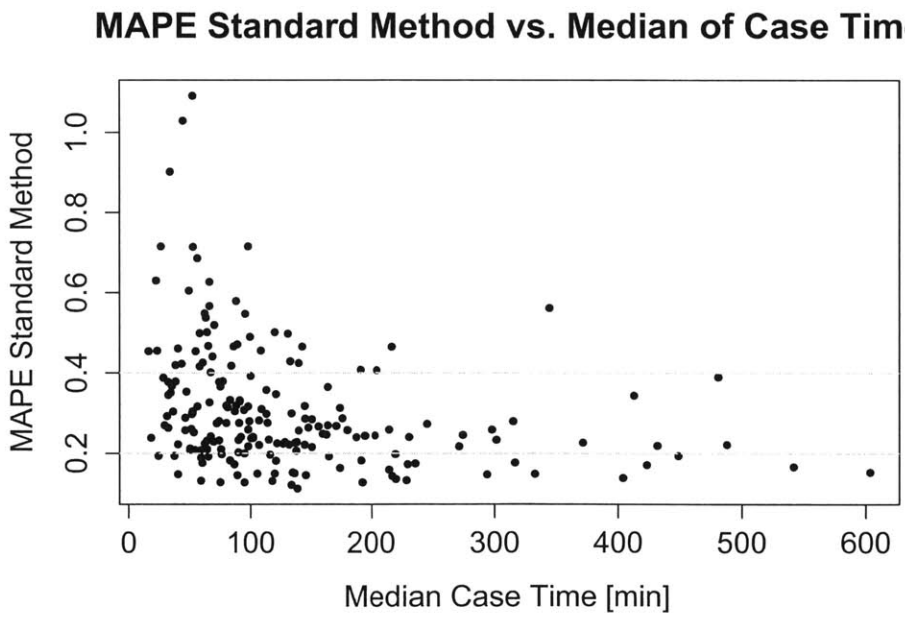


Note: Each point corresponds to a procedure type, and the vertical bars to the 95% bootstrap CI of the standard deviation and coefficient of variation, respectively.

Figure C-3: Performance of the standard method.



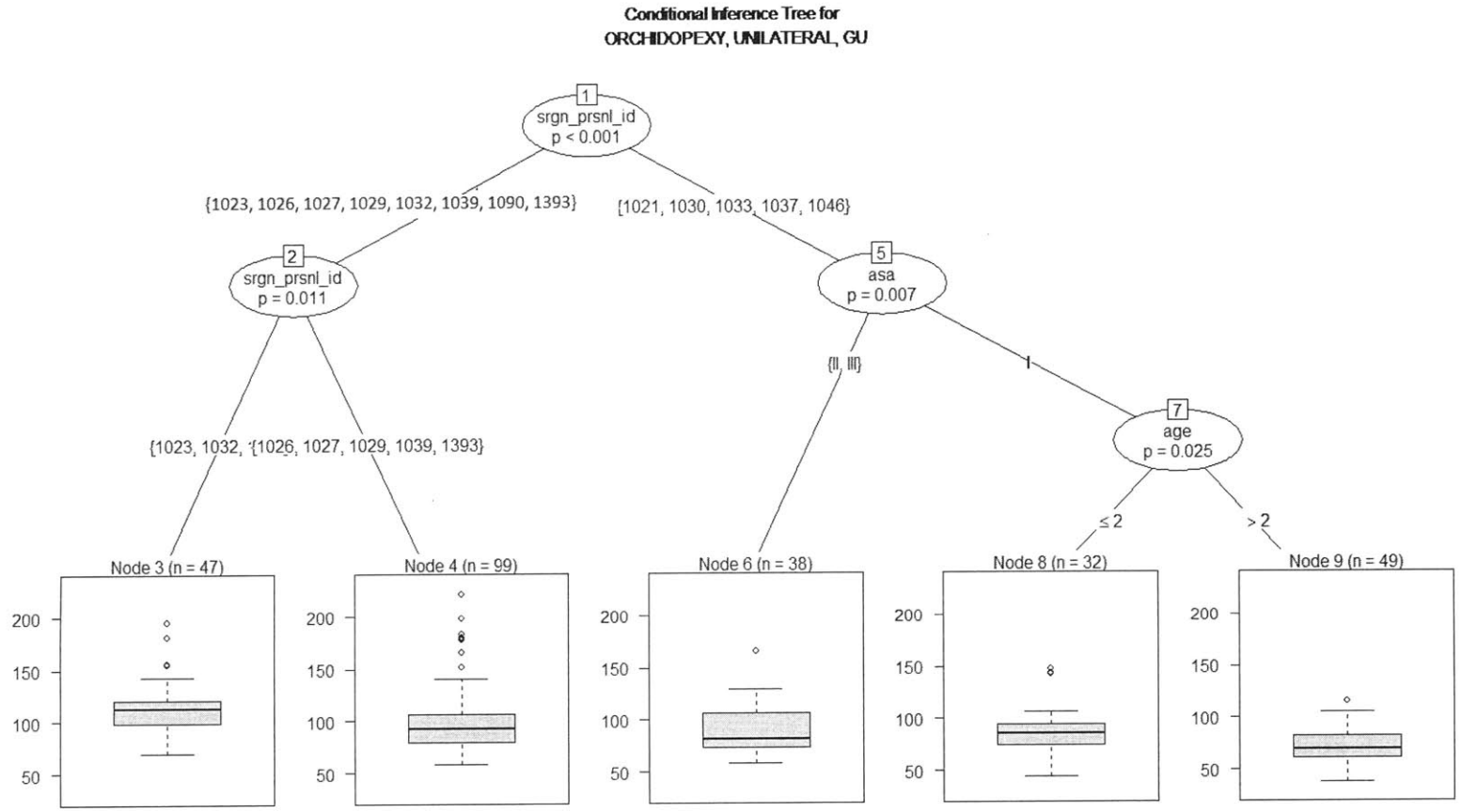
(a) MAE of the standard method vs. median case time. The slope of the regression line is 19.3 (15.6, 24.8)%, and the intercept is 10.2(5.32, 14.69) absolute minutes. The 95% CI linear regression and R-Squared were computed via bootstrapping (BCa, 1000 replications).



(b) MAPE of the standard method vs. median case time. The horizontal lines are markers at 20% and 40% MAPE.

Note: Each data point corresponds to a procedure type (195 different procedures).

Figure C-4: Example: conditional inference regression tree for Orchidopexy Unilateral.



Note: The first split is based on surgeon; the following split (node 2) in the left hand side is again based on surgeon resulting in terminal nodes 3 and 4. In the right hand side after the first surgeon split (node 5), the tree splits based on ASA index obtaining on the left daughter a terminal node (node 6) for the ASA levels II and III. The right daughter node (node 7) is further split based on patient's age resulting in two terminal nodes for patients of age two or less years (node 8) and older than 2 years (node 9). Each of the leaves contain train samples that are similar in terms of case time. The distribution of case time for observations in the final buckets is shown in the box plots.

Table C.1: Summary of case time statistics and RT results.

key	Procedure Type	N	Mean Case Time [min]	Sd. Case Time [min]	CV	Srgn	ICU	ASA	Weight	Age
1	TONSILLECTOMY/ADENOIDECTOMY	753	54.95(53.88, 56.12)	15.77(14.6, 17.29)	0.29	1	2		2	
2	TYMPANOSTOMY/TUBES	2086	29.24(28.37, 30.16)	21.13(19.88, 22.55)	0.72	3		1		2
3	CIRCUMCISION	994	61.96(61, 62.99)	15.85(14.89, 16.99)	0.26	1				
4	DIRECT LARYNGOSCOPY/BRONCHOS	468	65.88(62.49, 69.74)	38.78(34.43, 44.64)	0.59			1		
5	ARTHROSCOPY KNEE	718	71.65(69.84, 73.62)	25.68(23.6, 28.08)	0.36	1				
6	ADENOIDECTOMY	639	47.93(46.75, 49.19)	15.7(14.33, 17.41)	0.33	1				
10	ARTHROSCOPY KNEE ACL RECONST	287	114.87(111.9, 118.26)	27.48(24.35, 32.02)	0.24	1			2	2
12	ORCHIDOPEXY, UNILATERAL	410	93.25(90.71, 96.13)	28.19(25.63, 31.32)	0.3	1		2		3
13	PORT-A-CATH INSERTION	308	117.11(113.49, 121.09)	34.14(30.39, 39.46)	0.29	1				
14	HARDWARE REMOVAL BURIED PIN	318	75.26(71.64, 79.31)	34.61(30.99, 38.89)	0.46	1			2	
15	ESOPHAGOGASTRODUODENOSCOPY	332	69.91(66.35, 74.06)	36.13(31.85, 42)	0.52			2	1	
16	EXCISION LESION SMALL	337	77.41(74.17, 81.03)	32.02(29.32, 35.45)	0.41	1				
17	BOTOX INJECTION	325	30.42(28.98, 32.05)	13.87(12.45, 16.15)	0.46	1				
18	EXAM UNDER ANESTHESIA EARS	301	50.92(46.96, 55.63)	38.57(33.28, 46.31)	0.76			1		
22	COLONOSCOPY	282	103.07(99.19, 107.38)	35.44(31.66, 40.2)	0.34	1		2		
24	HERNIA REPAIR, INGUINAL/HYDROCEL	274	73.5(71.19, 76.18)	20.84(18.84, 23.96)	0.28	1				
25	MICRODIRECT LARYNGOSCOPY/BRONCHOS	174	63.36(58.4, 69.21)	36.02(31.43, 41.99)	0.57			1		2
26	PORT-A-CATH REMOVAL	260	64.6(62.48, 66.98)	18.22(16.28, 20.83)	0.28	1				
27	MEATOTOMY, URETHRAL	261	34.82(33.47, 36.33)	11.78(10.49, 13.44)	0.34	1				
28	ARTHROSCOPY HIP OSTEOPLASTY	132	124.83(119.7, 130.79)	32.16(27.52, 38.39)	0.26	1			2	
29	CYSTOSCOPY	238	69.56(65.36, 75.92)	40.97(31.74, 55.22)	0.59	1				
32	GROWING SPINAL RODS LENGTHENING	126	113.02(106.61, 122.47)	43.78(33.76, 57.08)	0.39	1				
35	CADD: DIRECT LARYNGOSCOPY/BRONCHOS	58	56(49.6, 63.72)	27.92(22.64, 35.85)	0.5		1			
38	EPIPHYSIODESIS	79	86.24(81.15, 91.67)	24.1(20.75, 28.55)	0.28	1				
41	STEROID INJECTION OF JOINT	171	33.43(31.82, 35.35)	11.8(10.21, 14.34)	0.35	1				2
43	ESOPHAGOGASTRODUODENOSCOPY	120	90.56(85.08, 97.18)	33.89(28.67, 41.15)	0.37					1
54	LYSIS OF PENILE ADHESIONS	136	50.63(47.21, 54.32)	21.01(18.71, 24.06)	0.41	1				
56	LUMBAR PUNCTURE	132	42.58(39.77, 45.79)	17.48(15.13, 20.53)	0.41				1	
61	HERNIA REPAIR, INGUINAL, UNILATERAL	110	82.18(77.46, 88.13)	28.79(23.35, 38.73)	0.35	1				
76	SHUNT, VP INSERTION/REVISION/REMOVAL	57	178.54(160.26, 202.11)	80.09(63.91, 101.09)	0.45				1	
79	ARTHROSCOPY HIP	55	120.8(112.27, 132.2)	37.19(26.97, 51.09)	0.31	1			2	
80	LAPAROTOMY, EXPLORATORY	67	365.48(326.23, 413.12)	182.22(150.52, 228.85)	0.5		1			
93	HARDWARE REMOVAL RADIUS/ULNA	89	57.17(52.49, 62.36)	23.68(20.41, 29.22)	0.41					1
94	TENOTOMY ACHILLES, PERCUTANEOUS	84	42.42(37.86, 48.89)	25.21(19.69, 32.3)	0.59				1	
100	ARTHROSCOPY KNEE WITH OPEN MEDIAL	71	78.52(73, 85.72)	26.94(21.39, 34.18)	0.34	1				
103	LACRIMAL DUCT PROBE, NO TUBES	78	29.59(26.68, 33.76)	15.55(11.65, 23.77)	0.53			1		
117	EXCISION LESION LARGE	45	122.07(103.38, 159.6)	87.28(52.4, 160.92)	0.71	1				
166	CHALAZION EXCISION	42	38.81(34.71, 43.36)	14.34(11.81, 18.04)	0.37	1				
200	CHORDEE RELEASE MILD	35	97.03(85.85, 110.16)	36.83(30, 49.92)	0.38	1				

Note: This table includes the set of 39 procedure types for which the RT identified correlation between case time and the studied factors. The notation $x(x^-, x^+)$ represents the point estimates and 95% confidence interval of the statistic. The last five columns correspond to the studied factors and the numbers indicate the order in which the RT branched. For example, for the first procedure type, the RT split on surgeon identity in the first iteration then, the resulting daughter nodes are split based on Patient's Weight and ICU request factors

Bibliography

[1] J. J. Baker and G. F. Boyd. Activity-based costing in the operating room at valley view hospital. *Journal of health care finance*, 24(1):1–9, 1996.

[2] J. T. Blake and M. W. Carter. A goal programming approach to strategic resource allocation in acute care hospitals. *European Journal of Operational Research*, 140(3):541–561, 2002.

[3] J. T. Blake, F. Dexter, and J. Donald. Operating room managers' use of integer programming for assigning block time to surgical groups: A case study. *Anesthesia & Analgesia*, 94(1):143–148, jan 2002.

[4] F. Bravo, M. Braun, V. Farias, and R. Levi. Optimization-driven framework to understand healthcare networks cost and resource allocation. *In preparation*, pages 1–33, 2015.

[5] F. Bravo, R. Levi, L. R. Ferrari, and M. L. McManus. The nature and sources of variability in pediatrics surgical case duration. *Pediatric Anesthesia (accepted manuscript)*, 2015.

[6] F. Bravo, R. Levi, G. Perakis, and G. Romero. A risk-sharing pricing contract in B2B service supply chains: an application to healthcare. *In preparation*, pages 1–42, 2015.

[7] A. Burnetas, S. M. Gilbert, and C. E. Smith. Quantity discounts in single-period supply contracts with asymmetric demand information. *IIE Transactions*, 39(5):465–479, 2007.

[8] G. P. Cachon. Handbooks in operations research and management science: supply chain management. *Supply chain coordination with contracts*, 2003.

[9] G. P. Cachon and M. a. Lariviere. Contracting to assure supply: How to share demand forecasts in a supply chain. *Management Science*, 47(5):629–646, 2001.

[10] G. P. Cachon and M. A. Lariviere. Supply chain coordination with revenue-sharing contracts: strengths and limitations. *Management Science*, 51(1):30–44, 2005.

[11] R. Caldentey and M. Haugh. Optimal control and hedging of operations in the presence of financial markets. *Mathematics of Operations Research*, 31(2):285–304, 2006.

[12] R. Capettini, C. W. Chow, and A. H. McNamee. On the need and opportunities for improving costing and cost management in healthcare organizations. *Managerial Finance*, 24(1):46–59, 1998.

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

[14] M. Chalkley and J. M. Malcomson. Contracting for health services when patient demand does not reflect quality. *Journal of health economics*, 17(1):1–19, 1998.

[15] M. Chalkley and J. M. Malcomson. Cost sharing in health service provision: An empirical assessment of cost savings. *Journal of Public Economics*, 84(2):219–249, 2002.

[16] C.-H. Chiu and T.-M. Choi. Supply chain risk analysis with mean-variance models: a technical review. *Annals of Operations Research*, pages 1–19, 2013.

[17] R. Cooper. Cost classification in unit-based and activity-based manufacturing cost systems. *Journal of Cost Management*, 4(3):4–14, 1990.

[18] R. Cooper and R. S. Kaplan. Profit priorities from activity-based costing. *Harvard Business Review*, 69(3):130–135, 1991.

[19] R. Cooper and R. S. Kaplan. Activity-based systems: Measuring the costs of resource usage. *Accounting Horizons*, 6(3):1–13, 1992.

[20] C. J. Corbett and G. a. DeCroix. Shared-savings contracts for indirect materials in supply chains: channel profits and environmental impacts. *Management Science*, 47(7):881–893, 2001.

[21] C. J. Corbett, G. a. Decroix, and A. Y. Ha. Optimal shared-savings contracts in supply chains: Linear contracts and double moral hazard. *European Journal of Operational Research*, 163(3):653–667, 2005.

[22] F. Dexter, E. U. Dexter, and J. Ledolter. Influence of procedure classification on process variability and parameter uncertainty of surgical case durations. *Anesthesia and Analgesia*, 110(4):1155–1163, 2010.

[23] F. Dexter, E. U. Dexter, and J. Ledolter. Importance of appropriately modeling procedure and duration in logistic regression studies of perioperative morbidity and mortality. *Anesthesia and Analgesia*, 113(5):1197–201, 2011.

[24] F. Dexter, E. U. Dexter, D. Masursky, and N. A. Nussmeier. Systematic review of general thoracic surgery articles to identify predictors of operating room case durations. *Anesthesia and Analgesia*, 106(4):1232–41, 2008.

[25] F. Dexter, R. H. Epstein, and J. Ledolter. Estimating surgical case durations and making comparisons among facilities: identifying facilities with lower anesthesia professional fees. *Anesthesia and Analgesia*, 116(5):1103–1115, 2013.

[26] F. Dexter, R. H. Epstein, J. D. Lee, and J. Ledolter. Automatic updating of times remaining in surgical cases using bayesian analysis of historical case duration data and "instant messaging" updates from anesthesia providers. *Anesthesia and Analgesia*, 108(3):929–940, 2009.

[27] F. Dexter, R. H. Epstein, R. D. Traub, and Y. Xiao. Making management decisions on the day of surgery based on operating room efficiency and patient waiting times. *Anesthesiology*, 101(6):1444–53, 2004.

[28] F. Dexter and J. Ledolter. Bayesian prediction bounds and comparisons of operating room times even for procedures with few or no historic data. *Anesthesiology*, 103(6):1259–167, 2005.

[29] F. Dexter and R. D. Traub. How to schedule elective surgical cases into specific operating rooms to maximize the efficiency of use of operating room time. *Anesthesia and Analgesia*, 94(4):933–42, 2002.

[30] Q. Ding, L. Dong, and P. Kouvelis. On the integration of production and financial hedging decisions in global markets. *Operations Research*, 55(3):470–489, 2007.

[31] R. J. Dolan. Quantity Discounts: Managerial issues and research opportunities. *Marketing Science*, 6(1):1–22, Feb. 1987.

[32] D. Dranove and M. Shanley. Cost reductions or reputation enhancement as motives for mergers: The logic of multihospital systems. *Strategic Management Journal*, 16(1):55–74, 1995.

[33] M. J. Eijkemans, V. H. Mark, T. Nguyen, E. Boersma, E. W. Steyerberg, and G. Kazemier. Predicting the unpredictable: A new prediction model for operating room times using Individual characteristics and the surgeon's estimate. *Anesthesiology*, 112(1):41–49, 2010.

[34] R. P. Ellis and T. G. McGuire. Provider behavior under prospective reimbursement: Cost sharing and supply. *Journal of health economics*, 5(2):129–151, 1986.

[35] Engelberg Center for Health Care Reform. Aco toolkit, learning network. Technical report, 2011.

[36] J. J. Escarce, J. A. Shea, and W. Chen. Segmentation of hospital markets: where do hmo enrollees get care? *Health Affairs*, 16(6):181–192, 1997.

[37] M. Esmaeili, M.-B. Aryanezhad, and P. Zeephongsekul. A game theory approach in seller-buyer supply chain. *European Journal of Operational Research*, 195(2):442–448, 2009.

[38] R. Feldman, J. Kralewski, J. Shapiro, and H.-C. Chan. Contracts between hospitals and health maintenance organizations. *Health Care Management Review*, 15(1):47–60, 1990.

[39] S. A. Finkler. *Essentials of cost accounting for health care organizations.* Aspen Publishers, 1994.

[40] P. C. Fuloria and S. A. Zenios. Outcomes-adjusted reimbursement in a healthcare delivery System. *Management Science*, 47(6):735–751, 2001.

[41] V. Gaur and S. Seshadri. Hedging inventory risk through market instruments. *Manufacturing & Service Operations Management*, 7(2):103–120, 2005.

[42] F. Gorunescu, S. I. McClean, and P. H. Millard. A queueing model for bed-occupancy management and planning of hospitals. *Journal of the Operational Research Society*, 53(1):19–24, 2002.

[43] F. Harinck, E. Van Dijk, I. Van Beest, and P. Mersmann. When gains loom larger than losses: reversed loss aversion for small amounts of money. *Psychological science*, 18(12):1099–105, 2007.

[44] P. R. Harper and A. K. Shahani. Modeling for the planning and management of bed capacities in hospitals. *Journal of the Operational Research Society*, 53(1):11–18, 2002.

[45] A. C. Hax and D. Candea. *Production and inventory management.* Prentice-Hall, Englewood Cliffs, NJ, 1984.

[46] B. He, F. Dexter, A. Macario, and S. A. Zenios. The timing of staffing decisions in hospital operating rooms: incorporating workload heterogeneity into the newsvendor problem. *Manu & Service Operations & Service Operations*, 14(1):99–114, 2012.

[47] B. Hezarkhani and W. Kubiak. Coordinating contracts in scm: a review of methods and literature. *Decision Making in Manufacturing and Services*, 4(1):5–28, 2010.

[48] T. Hothorn, K. Hornik, and A. Zeileis. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15(3), 2006.

[49] W. L. Hughes and S. Y. Soliman. Short-term case mix management with linear programming. *Hospital & health services administration*, 30(1):52–60, 1984.

[50] E. P. Jack and T. L. Powers. A review and synthesis of demand management, capacity management and performance in health-care services. *International Journal of Management Reviews*, 11(2):149–174, 2009.

[51] W. Jack. Purchasing health care services from providers with unknown altruism. *Journal of Health Economics*, 24(1):73–93, 2005.

[52] N. Jain, S. Hasija, and D. G. Popescu. Optimal contracts for outsourcing of repair and restoration services. *Operations Research*, 61(6):1295–1311, 2013.

[53] W. Jammernegg and P. Kischka. Newsvendor problems with var and cvar consideration. In *Handbook of Newsvendor Problems*, pages 197–216. Springer, 2012.

[54] H. Jiang, Z. Pang, and S. Savin. Performance-based contracts for outpatient medical services. *Manufacturing & Service Operations Management*, 14(4):654–669, 2012.

[55] D. Kahneman and A. Tversky. Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society*, pages 263–291, 1979.

[56] R. K. Kanter and F. Dexter. Criteria for identification of comprehensive pediatric hospitals and referral regions. *The Journal of Pediatrics*, 146(1):26–9, 2005.

[57] E. P. Kao and G. G. Tung. Bed allocation in a public health care delivery system. *Management Science*, 27(5):507–520, 1981.

[58] R. S. Kaplan and M. E. Porter. How to solve the cost crisis in health care. *Harv Bus Rev*, 89(9):46–52, 2011.

[59] R. S. Kaplan and M. L. Witkowski. Better accounting transforms health care delivery. *Accounting Horizons*, 28(2):365–383, 2014.

[60] A. A. Khaliq and S. L. Walston. Which hospitals give discounts? The role of institutional and environmental factors. *Hospital topics*, 85(2):13–8, Jan. 2007.

[61] R. A. Krishnan, S. Joshi, and H. Krishnan. The influence of mergers on firms' product-mix strategies. *Strategic Management Journal*, 25(6):587–611, 2004.

[62] P. Krokhmal, J. Palmquist, and S. Uryasev. Portfolio optimization with conditional value-at-risk objective and constraints. *Journal of Risk*, 4:43–68, 2002.

[63] M. A. Lariviere and E. L. Porteus. Selling to the newsvendor: An analysis of price-only contracts. *Manufacturing & service operations management*, 3(4):293–305, 2001.

[64] F. J. Lexa, T. Mehta, and A. Seidmann. Managerial accounting applications in radiology. *Journal of the American College of Radiology*, 2(3):262–270, 2005.

[65] M. Lu and C. Donaldson. Performance-based contracts and provider efficiency: The state of the art. *Disease Management and Health Outcomes*, 7(3):127–137, 2000.

[66] G. Ma and E. Demeulemeester. A multilevel integrative approach to hospital case mix and capacity planning. *Computers & Operations Research*, 40(9):2198–2207, 2013.

[67] R. Mannion, G. Marini, and A. Street. Implementing payment by results in the English NHS: changing incentives and the role of information. *Journal of health organization and management*, 22(1):79–88, Jan. 2008.

[68] M. Miraldo, L. Siciliani, and A. Street. Price adjustment in the hospital sector. *Journal of Health Economics*, 30(1):112–125, 2011.

[69] Moody's Investor Service. For-profit investments in non-for-profit hospitals signals more consolidation ahead, 2010.

[70] J. P. Newhouse. *Pricing the priceless: a health care conundrum.* MIT Press, 2002.

[71] L. O'Neill and F. Dexter. Tactical increases in operating room block time based on financial data and market growth estimates from data envelopment analysis. *Anesthesia and analgesia*, 104(2):355–68, 2007.

[72] G. Perakis and G. Roels. The price of anarchy in supply chains: Quantifying the efficiency of price-only contracts. *Management Science*, 53(8):1249–1268, 2007.

[73] J. Rice. *Mathematical Statistics and Data Analysis.* Cengage Learning, 2006.

[74] W. Robbins and N. Tuntiwongpiboom. Linear programming a useful tool in case-mix management. *Healthcare financial management: journal of the Healthcare Financial Management Association*, 43(6):114–116, 1989.

[75] R. T. Rockafellar and S. Uryasev. Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42, 2000.

[76] G. Roels, U. S. Karmarkar, and S. Carr. Contracting for collaborative services. *Management Science*, 56(5):849–863, May 2010.

[77] W. Samuelson and R. Zeckhauser. Status quo bias in decision making. *Journal of risk and uncertainty*, 1:7–59, 1988.

[78] M. Shaked and J. G. Shanthikumar. *Stochastic orders.* Springer, 2007.

[79] J. H. Silber, P. R. Rosenbaum, X. Zhang, and O. Even-Shoshan. Influence of patient and hospital characteristics on anesthesia time in medicare patients undergoing general and orthopedic surgery. *Anesthesiology*, 106(2):356–64, 2007.

[80] K. Singh and M. Xie. Bootstrap: A statistical method. 2010.

[81] B. Smallman and F. Dexter. Optimizing the arrival, waiting, and NPO times of children on the day of pediatric endoscopy procedures. *Anesthesia and Analgesia*, 110(3):879–87, 2010.

[82] V. L. Smith-Daniels, S. B. Schweikhart, and D. E. Smith-Daniels. Capacity management in health care services: Review and future research directions. *Decision Sciences*, 19(4):889–919, 1988.

[83] W. E. Spangler, J. H. May, D. P. Strum, and L. G. Vargas. A data mining approach to characterizing medical code usage patterns. *Journal of Medical Systems*, 26(3):255–75, 2002.

[84] D. Steinberg and P. Colla. CART: classification and regression trees. In X. Wu and V. Kumar, editors, *The Top Ten Algorithms in Data Mining*, chapter 10, page 179. CRC Press, 2009.

[85] P. S. Stepaniak, C. Heij, and G. de Vries. Modeling and prediction of surgical procedure times. *Statistica Neerlandica*, 64(1):1–18, 2010.

[86] P. S. Stepaniak, C. Heij, G. H. H. Mannaerts, M. de Quelerij, and G. de Vries. Modeling procedure and surgical times for current procedural terminology-anesthesia-surgeon combinations and evaluation in terms of case-duration prediction and operating room efficiency: a multicenter study. *Anesthesia and Analgesia*, 109(4):1232–45, 2009.

[87] D. P. Strum, J. H. May, A. R. Sampson, L. G. Vargas, and W. E. Spangler. Estimating times of surgeries with two component procedures: comparison of the lognormal and normal models. *Anesthesiology*, 98(1):232–240, 2003.

[88] D. P. Strum, J. H. May, and L. G. Vargas. Modeling the uncertainty of surgical procedure times: comparison of log-normal and normal models. *Anesthesiology*, 92(4):1160–1167, 2000.

[89] D. P. Strum, A. R. Sampson, J. H. May, and L. G. Vargas. Surgeon and type of anesthesia predict variability in surgical procedure times. *Anesthesiology*, 92(5):1454–1466, 2000.

[90] K. T. Talluri and G. J. Ryzin. *The theory and practice of revenue management*, volume 68. springer, 2006.

[91] F.-S. Tseng and Y. Yeh. Maintenance outsourcing coordination with risk-averse contractors. *Journal of the Operational Research Society*, 65(11):1–10, 2013.

[92] A. Tversky and D. Kahneman. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty*, 5(4):297–323, 1992.

[93] S. Udpa. Activity-based costing for hospitals. *Health Care Management Review*, 21(3):83, 1996.

[94] S. Viswanathan and Q. Wang. Discount pricing decisions in distribution channels with price-sensitive demand. *European Journal of Operational Research*, 149(3):571–587, Sept. 2003.

[95] J. Von Neumann and O. Morgenstern. Theory of games and economic behavior. *Bull. Amer. Math. Soc*, 51(7):498–504, 1945.

[96] R. E. Wachtel and F. Dexter. Differentiating among hospitals performing physiologically complex operative procedures in the elderly. *Anesthesiology*, 100(6):1552–61, 2004.

[97] R. E. Wachtel and F. Dexter. Tactical increases in operating room block time for capacity planning should not be based on utilization. *Anesthesia and analgesia*, 106(1):215–26, 2008.

[98] C. X. Wang, S. Webster, and S. Zhang. Newsvendor models with alternative risk preferences within expected utility theory and prospect theory frameworks. In *Handbook of newsvendor problems*, pages 177–196. Springer, 2012.

[99] K. Weng. Modeling quantity discounts under general price-sensitive demand functions: Optimal policies and relationships. *European Journal of Operational Research*, 86(2):300–314, 1995.

[100] I. Wright, C. Kooperberg, B. Bonar, and G. Bashein. Statistical Modeling to Predict Elective Surgery. *Anesthesiology*, (85):1235–1245, 1996.

[101] L. Yang, M. Xu, G. Yu, and H. Zhang. Supply chain coordination with cvar criterion. *Asia-Pacific Journal of Operational Research*, 26(01):135–160, 2009.

[102] G. J. Young, J. E. Burgess, and D. Valley. Competition among hospitals for HMO business: effect of price and nonprice attributes. *Health services research*, 37(5):1267–89, 2002.

[103] J. P. Young. *A queuing theory approach to the control of hospital inpatient census.* John Hopkins University, 1963.

[104] D. Zhang, H. Xu, and Y. Wu. Single and multi-period optimal inventory control models with risk-averse constraints. *European Journal of Operational Research*, 199(2):420–434, 2009.

[105] J. Zhou and F. Dexter. Method to assist in the scheduling of add-on surgical dases-upper prediction bounds for surgical case durations based on the dog-normal distribution. *Anesthesiology*, 89(5):1228–1232, 1998.