Incentives in U.S. Healthcare Operations

Tinglong Dai

Submitted to Tepper School of Business, Carnegie Mellon University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy in Management of Manufacturing and Automation

Dissertation Committee:

Professor Sridhar Tayur (Co-Chair)

Professor Katia Sycara (Co-Chair)

Professor Mustafa Akan

Professor Soo-Haeng Cho

Professor Kinshuk Jerath

Professor R. Ravi

April 2013

Abstract

In my dissertation, I aim to understand incentives in U.S. healthcare operations based on my collaboration with various health organizations.

In my first essay, I investigate the underlying operational and economic drives behind physicians' test-ordering behavior in an outpatient setting, motivated by a collaborative study with University of Pittsburgh Medical Center (UPMC) Eye Center. I model the physician-patient interaction under the strategic queueing framework, and show that insurance coverage is a key driving force of overtesting. Our further analysis reveals that simply expanding cost-sharing does not constitute the solution: (i) While existing studies hold that lower out-of-pocket expenses lead to higher consumption levels, we refine this statement by showing that the copayment and the coinsurance rate drive the consumption toward different directions. (ii) Setting a low reimbursement ceiling alone cannot eliminate overtesting. (iii) The joint effect of misdiagnosis concerns and insurance coverage can lead to both overtesting and undertesting even when there is no reimbursement ceiling. These and other results continue to hold under more general conditions and so are robust. We also consider other extensions, including patient heterogeneity and information asymmetry in physician type.

Motivated by the influenza vaccine industry, in my second essay, I study a supply chain contracting problem under the presence of uncertainties that are related to product design, delivery, and demand, respectively. The supply chain consists of a manufacturer and a retailer, where the retailer places an order before the flu season starts and the manufacturer decides when to produce the products. Because production after the design freeze can result in late deliveries and hence lost sales, the manufacturer may initiate production prior to the design freeze at its own risk. I show that a negative feedback loop in the firms' incentives may arise in this supply chain; as a result, some of the traditional coordinating contracts (e.g., revenue sharing) could perform even worse than a wholesale price contract. To break the negative feedback loop requires complex contracts that are reported in practice but never studied in the literature. In view of the complexity of the coordinating contracts, I also analyze two simpler formats and show that they are efficient in chain coordination under various settings.

My third essay applies queueing and game theories to model a proposed organ donation policy. I model the current organ donation and allocation system, and evaluate the effect of introducing the donor priority policy under which registered organ donors are given priority to receive organs over non-donors, a frequently discussed policy being considered by U.S. policy makers. I characterize the equilibrium donating behavior and show that, as opposed to popular beliefs and extant literature, the social welfare can be worse off after introducing the donor priority policy due to the unbalanced incentive structure for individuals with heterogeneous health status. Finally, I propose a simple freeze-period mechanism and prove that it improves the welfare outcome of the donor priority rule by increasing the donation rate without distorting the quality distribution of the donated organs.

Acknowledgements

Everyday at Carnegie Mellon, I have no doubts that I belong here. No school ever gave me a stronger sense of belonging than Carnegie Mellon does. I was able to enjoy my time here mostly because of my advisors, Prof. Katia Sycara and Prof. Sridhar Tayur, who are "the really great [who] make you feel that you, too, can become great" (Mark Twain). They have permanently changed my life. Katia admitted me to this unique PhD program six year ago, and has been extremely generous in supporting me and helping me grow both professionally and personally. Sridhar brought me to the field of Healthcare Operations Management in the spring of 2008, and has since been the greatest source of motivation behind my healthcare research. I am forever grateful to Katia and Sridhar for their trust, encouragement, and tolerance in unmeasurable proportions.

I felt incredibly fortunate to have worked closely with Prof. Mustafa Akan, Prof. Soo-Haeng Cho, Prof. Kinshuk Jerath, and Prof. Fuqiang Zhang. There are so many fascinating things to say about each of them, but to put it radically simply, Mustafa taught me about grace, Soo-Haeng challenged me to present things clearly, Kinshuk showed me the power of thinking and articulation, and Fuqiang refined my taste in identifying research directions. My thanks also go to Prof. R. Ravi, who kindly served in my dissertation committee.

I benefited tremendously from a number of medical doctors: Dr. Joshua Rheinbolt and Dr. Robert Noecker (both at University of Pittsburgh Medical Center), Dr. Heidi Yeh and Dr. James Markmann (both at Massachusetts General Hospital in Cambridge, MA), and, of course, my wife Dr. Ruizhi Wang.

Several faculty members provided me with invaluable guidance at various stages: Prof. Kannan Srinivasan, Prof. Alan Scheller-Wolf, Prof. Laurens Debo (Chicago Booth), Prof. Baris Ata, Prof. Mor Harchol-Balter, Prof. Bahar Biller, Prof. Michael Lewis (University of Pittsburgh), Prof. Judith Lave (University of Pittsburgh), Prof. Laurie Weingart, Prof. Nicola Secomandi, Prof. Sunder Kekre, Prof. Gérard Cornuéjols, Prof. Baohong Sun, Prof. Rachna Shah (Minnesota), Prof. Nagesh Gavirneni (Cornell), Prof. Jeff Hong (HKUST), and Prof. Nicholas Hall (OSU).

I appreciate the Accounting faculty for treating me as a part of the close-knit group. Prof. Yuji Ijiri, an Accounting pioneer, is an incredibly humble mentor with whom I had many memorable conversations. Prof. Jonathan Glover's seminar course was a major source of my research ideas; every hallway conversation with Jon imbued me with energy. Prof. Pierre Jinghong Liang devoted a vast amount of time to consolidating my teaching profile and setting me in the right perspective of our profession. Prof. Jing Li trusted me so much that she offered me a recitation leadership job at a time when I knew nothing about Financial Accounting and had never taught any class before.

Prof. Natalie Baker-Shirer at the Drama School is a transforming teacher, expanding my curiosity for poetry and the arts while elevating my speaking skills.

I learned a lot from being a TA for faculty members in various disciplines, including Prof. Lloyd Corder, Prof. David Tungate, Prof. Jay Apt, the late Prof. Lester Lave, Prof. Anisha Ghosh, and Prof. Michael Trick. They have had profound impacts on me.

Thanks to our beloved Mr. Lawrence Rapp for making Tepper such a fabulous place. Thanks to Mrs. Marliese Bonk at the Robotics Institute for her professionalism and friendship.

Four professors paved the way for my PhD studies. Prof. Xiangtong Qi (HKUST) introduced me to the world of Operations Research in 2004 and has always been supportive of my further studies. Prof. Zhaolin Li (University of Sydney) gave me an initial appreciation of managerial insights during 2006–2007. Prof. Candace Yano (University of California, Berkeley) motivated me to do truly new things and avoid following others during her visit to HKUST in 2005. Prof. Rodney Parker (Chicago Booth, then at Yale) strongly recommended that I join Tepper's PhD program and work with Prof. Sridhar Tayur, adding that "Carnegie Mellon has a great history of graduating terrific students in a very intellectually vibrant environment."

Numerous friends made my PhD studies an interesting period of time. The following list is in no particular order and definitely incomplete: Min Cao, Shannon Ma, Michele Duffala, Zhaleh Semnani-Azad, Hong Qu, Wei Wang, Tim Allen, Marrick Smith, Joan Bruni, Deanna Ding, Jian Ni, Meng Zhu, Ming Jin, Qi Fei, Bin Fu, Liu Liu, Guang Xiang, Xiong Zhang, Xi Chen, Bin Fan, Shuang Su, Yingze Wang, Minjie Qian, Yingda Lu, Zhan

Zhang, Vivian Lui, Shiwei Qi, Emre Nadar, Ronghuo Zheng, Ying Xu, Hao Xue, Vishal Ahuja, Lin Xiao, Cong Wang, Baojun Jiang, Gaoqing Zhang, Xin Fang, Vincent Slaugh, Victor Sanchez-Anguix, Nilanjan Chakraborty, Lingzhi Luo, Hanrui Zhang, Shixuan Sun, Zhiyu Xu, Zhou Xu, Yanzhi Li, Chen Lin, Nazli Turan, Cynthia Cheung, Jacqueline del Castillo, and William Wang.

Finally, and most importantly, I could never imagine anything without the love and support of my parents, my siblings, my late grandparents, my other family members, my parents and sister-in-law, my high school teacher Ligong Zhang, and my undergraduate philosophy professor Jiaqi Chen. In particular, my deepest thanks go to my wife Ruizhi, who never ceased to believe in and support me with unconditional, everlasting love. This dissertation is dedicated to Ruizhi and our soon to be born baby.

To Ruizhi, and our soon to be born baby.

Contents

List of Tables

List of Figures

Introduction

Few would deny that the U.S. healthcare sector is experiencing a vexing crisis, as Berwick and Hackbarth (2012) acknowledge, “No matter how polarized politics in the United States have become, nearly everyone agrees that health care costs are unsustainable.” But the seemingly uncontainable cost growth is not the only worrying issue. The U.S. healthcare system also suffers from quality and access issues. It is no wonder that we live in an era with many discussions about how to reform our healthcare system, as Clayton Christensen et al. (2009) describe in *The Innovator’s Prescription*,

> “Almost every day somewhere in the United States, a group of health-care reformers convenes a conference. We’ve attended many of these. Almost without exception the participants talk past each other... They talk past one another because they don’t share a common language and a common understanding of the root causes of these problems.”

One crucial part of such root causes, I believe, is the complex web of incentives in the U.S. healthcare delivery system. In a hospital setting, Janet Currie observes,

> “The most striking thing [is] how rapidly the hospitals responded and how much they can change their service mix to try and attract the type of patients that are profitable. It doesn’t really make very much difference whether they’re private hospitals or public hospitals or for profit or not” (Clement 2012).

The literature of Healthcare Operations Management has flourished over the past decade as an increasingly influential part of the contemporary healthcare discourse. Most of the Healthcare OM literature, however, focuses on either OR applications to healthcare scenarios, or empirical studies of healthcare operations. Little attention has been paid to incentives in healthcare operations. On the other hand, there is a paucity of operations-level modeling in the Healthcare Economics literature. While benefiting from the two streams of literature, I was intrigued to study the incentives in U.S. healthcare operations, and found that the "GSIA approach" (Augier and March 2011), which unifies theoretic rigorousness and practical relevance, to be particularly suitable for my research.

My journey of healthcare research started with a collaborative project launched a few months after I became a PhD student. Together with Prof. Sridhar Tayur and two physicians, Dr. Joshua Rheinbolt and Dr. Robert Noecker, I aimed to help the University of Pittsburgh Medical Center (UPMC) Eye Center redesign its ocular imaging room, which had suffered from both long waiting times and underutilization. To uncover the problem underlying the puzzling symptoms, we videotaped the imaging room for two consecutive weeks, before spending half a year analyzing the patient flow through the network of ocular imaging devices. I subsequently conducted simulation-based optimization and proposed simple scheduling rules that helped boost the efficiency of movements in the imaging room, especially the reduction in waiting time without sacrificing the system throughput.

However, I soon realized that our efficiency-improving efforts for the imaging room were severely constrained by something beyond that room—physicians' orders for diagnostic tests. This led me into examining incentives behind physicians' test-ordering behavior in the outpatient setting (see Chapter 1 "Imaging Room and Beyond: The Underlying Economics behind Physicians' Test-Ordering Behavior in Outpatient Services"), including insurance coverage and risk of malpractice suits, as well as the impact of asymmetric information on physicians' heterogeneous skill levels. I uncovered many unexpected but theoretically robust and empirically verifiable insights about physician decision-making. For example, our analytical framework helps explain why overtesting (i.e., ordering unnecessary tests and procedures) could occur even when the payment system is non-fee-for-service, why limiting the reimbursable amount from the health payers would not eliminate overtesting, and why requiring physician service fees to be transparent to the public could worsen overtesting.

I gained from the UPMC experience a perspective on conducting research in healthcare operations: behind a seemingly conventional operational problem, there could be an intriguing web of incentives. This idea has influenced my research ever since. In another research project on the U.S. influenza vaccine supply chain (see Chapter 2 "Contracting for On-Time Delivery in the U.S. Influenza Vaccine Supply Chain"), I started with the following puzzle observed in the U.S. vaccine market: In many flu seasons, a shortage of influenza vaccine can occur even when the total supply is abundant. The catch, it turned out, was that late delivery of vaccine can cause significant reduction in demand due to cancellation of many vaccination campaigns. I found that the suboptimal on-time delivery of influenza vaccine is not a scheduling problem as one might expect. Rather, it involves a negative feedback loop between vaccine manufacturers and providers: On one hand, the manufacturer lacks the motivation to improve the on-time delivery performance, which leads to potential shrinkage of the market size; on the other hand, the demand loss incentivizes the retailer to order a low quantity, which further discourages the manufacturer's delivery-improving effort. The research shows that due to this negative feedback loop, several well-studied contracts can lead to severe suboptimal performance, and, in particular, the revenue-sharing contract can be inferior to the wholesale price contract. Breaking this loop thus becomes the key to improving the on-time delivery performance of the supply chain. The same modeling framework applies to a wide range of industries with uncertainties in delivery, demand, and design, including the fashion industry, where the trend is ever-shifting and many firms rely on suppliers far away from their points of sales.

I have also been working on the U.S. organ transplantation system. Specifically, I am interested in studying the incentives on the supply side—organ donation (see Chapter 3 "The Welfare Consequences of the Donor Priority Rule"). The current system is not effectively converting the public's support for organ donation into actual donation rates. I focus on analyzing the impact of introducing the donor priority rule, which endows previously registered organ donors with the priority for receiving organs when they are in need of cadaveric organs. My analysis has revealed a hidden cost of the donor priority rule: although the donor priority rule invariably increases the size of the organ donor registry, the pool of donated organs can be distorted and the overall social welfare can hence be worse off after adopting the donor priority rule if the population is differentially healthy. I then propose a simple freeze-period remedy that can increase the donation rate without distorting the quality distribution.

In summary, this dissertation reflects my healthcare research during my PhD studies, based on my collaboration with physicians, hospitals, pharmaceutical firms, and other healthcare organizations, and empowered by OR and Economics tools. I found my collaboration with practitioners essential to my research because from doing so, it became immediately evident which problems are the most pressing, important, and interesting. The technical foundation in OR and Economics, on the other hand, made it possible to analyze various operational issues that would otherwise be hard to capture. I hope these studies help achieve a better understanding of incentives in the U.S. health care delivery system.

Chapter 1

Imaging Room and Beyond: The Underlying Economics behind Physicians' Test-Ordering Behavior in Outpatient Services

Every time you walk into a doctor's office, it's implicit that someone else will be paying most or all of your bill; for most of us, that means we give less attention to prices for medical services than we do to prices for anything else. Most physicians, meanwhile, benefit financially from ordering diagnostic tests, doing procedures, and scheduling follow-up appointments. Combine these two features of the system with a third—the informational advantage that extensive training has given physicians over their patients, and the authority that advantage confers—and you have a system where physicians can, to some extent, generate demand at will.

"How American Health Care Killed My Father" by David Goldhill, in *The Atlantic*, September 2009

1.1 Background

Over the last few decades, a strong consensus has emerged among patients, physicians, and policy makers that health care is not delivered efficiently in the United States. One major aspect of the inefficiency in the healthcare system is the prescription of unnecessary diagnostic tests and medical procedures by physicians (hereafter referred to as "overtesting"). It was estimated that up to $765 billion per year, or 6 percent of the nation's GDP, is spent on unnecessary tests and treatments (Institute of Medicine 2010). Unfortunately, conventional cost containment strategies are not effective because "managers make resource allocation decisions, but doctors decide what the hospital does with those resources. A horizontal cleavage divides the clinical workers from the containment sector, and there is little cooperation between the two" (Carter 2002). To curb the phenomenon of overtesting, the first step is to understand its underlying drivers (Rao and Levin 2012).

So far, the most commonly cited explanation for overtesting is misaligned monetary incentives as a result of the fee-for-service payment system that gives physicians more revenue as they order more tests (Welch 2012). This is manifested in President Barack Obama's description of the health care industry as "a system of incentives where the more tests and services are provided, the more money we pay, ... a model that rewards the quantity of care rather than the quality of care; [a model] that pushes you, the doctor, to see more and more patients even if you can't spend much time with each, ... a model that has taken the pursuit of medicine from a profession—a calling—to a business" (White House 2009). Another popular explanation for overtesting is that physicians use excessive testing to ease their own concerns for misdiagnosis (Rao and Levin 2012).

Our collaborative study with the University of Pittsburgh Medical Center (UPMC) Eye Center, one of the top ophthalmology programs in the U.S., revealed a strikingly different picture. In the existing payment model at UPMC, insurance plans only approve payment for one test per day/per patient. Moreover, depending on the type of test and disease, insurance firms limit the number of reimbursable tests per year. For instance, when a physician orders three tests for a patient, the physician understands that only one test will be reimbursed by the insurance firm, and that the other two will not generate additional revenue. However, overtesting has been consistently observed at UPMC Eye Center despite physicians' general lack of direct monetary incentives and misdiagnosis concerns (Rheinbolt and Noecker 2009; Rheinbolt 2012).

Based on our interviews with physicians and patients at UPMC Eye Center, we identified three crucial factors behind patients' decisions to visit doctors' offices: out-of-pocket expense, waiting time, and service quality. First, in the U.S. healthcare market, the majority of patients are insured and pay less than the actual service charge. Second, long service queues influence patients' experiences to such an extent that patients desire monetary compensation for long waiting times (Alderman 2011), and waiting-time-tracking websites like `www.medwaittime.com` have emerged. Third, patients are concerned about the service quality, which is closely tied to the quantity of diagnostic tests, though the marginal return from ordering additional tests is diminishing (Mold et al. 2010). These three aspects echo the iron triangle of U.S. healthcare, namely, cost, access, and quality, as proposed by Kissick (1994).

To examine physicians' test-ordering behavior, our model captures key financial, operational, and clinical incentives that govern the interactions between the physician and patients. While the physician strikes a balance between system throughput and diagnostic certainty, patients optimally trade off between waiting time, out-of-pocket expense, and service quality. We characterize the physician's optimal service parameters and patients' queue joining decisions, which we refer to as the *market equilibrium*, as opposed to the *social optimum* in which the social welfare is maximized. The measure of inefficiency is the loss of social welfare with respect to the socially efficient administration of imaging tests. Our model reveals several interesting results that help understand the incentives behind physicians' test-ordering behavior.

First, we show that even in the absence of fee-for-service system and other commonly cited reasons, overtesting can still occur due to distortion caused by the insurance coverage that distorts the price signal. This is aligned with the literature has a different implication due to the analysis in the following.

Second, while existing studies hold that lower out-of-pocket expenses lead to higher consumption levels, we refine this statement by showing that the copayment and the coinsurance rate can drive the consumption toward *reverse* directions.

Third, when insurance firms impose reimbursement ceilings to physician practices, it essentially restricts physicians' pricing power. Under a reimbursement celing, we show that overtesting can nonetheless occur even under when the ceiling is low, and increasing the share of patients' cost-sharing (e.g., coinsurance) can induce more tests to be ordered.

Fourth, there are situations where physicians bear the risk of misdiagnosis that can be attributed to either misinterpretation of results from diagnostic tests, or failure to order adequate diagnostic tests. Contrary to conventional wisdom, we show that both overtesting and undertesting are possible outcomes with the introduction of misdiagnosis concerns. The underlying intuition is that physicians' misdiagnosis concerns push up the socially efficient consumption level..

In addition, we consider two extensions on patient heterogeneity and information asymmetry, respectively. We model patient heterogeneity in both insurance coverage and waiting costs, and show that such heterogeneity leads to more severe overtesting. We model information asymmetry in physicians' skill level, which helps reveal the effect of price transparency as well as technological advancements on physicians' test-ordering behavior.

1.1.1 Salient Features of UPMC Eye Center Scenario

We describe the salient features in ordering and conducting the imaging tests at UPMC Eye Center, which is representative of many academic, elective outpatient settings. Figure 1.1 shows the flow schematic of conducting imaging tests.

At UPMC Eye Center, when physicians order imaging tests, they typically order multiple tests all at once; this is in contrast to a sequential testing process (i.e., start from one test, and then decide whether to order further tests depending the information collected from the first test, and so on) in other clinical settings. A typical order can be a combination of tests such as OCT (which provides cross-sectional analysis of birefringent tissues

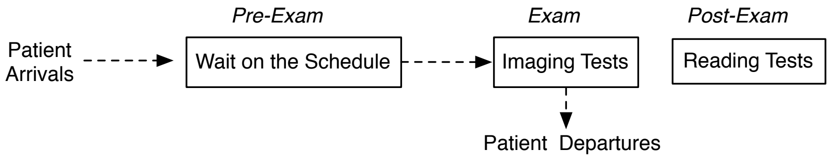


Figure 1.1: Flow schematic for imaging services

Note: This above schematic is adapted from Hopp and Lovejoy (2012).

in the eye), GDx (which measures the thickness of the retinal nerve to determine the occurrence of structural damage) and HRT3 (which provides a 3-D topography image of the optic nerve). Very few patients require a second batch of tests to be ordered for the same complaint.

As previously mentioned, the payment system is not fee for service, but the phenomenon of overtesting has been observed at UPMC Eye Center, as demonstrated by a significant variation in test-ordering patterns even for patients with comparable conditions (Dai et al. 2013). Due to extensive testing for patients, there is a high utilization rate that leads to patients' long waiting times for imaging tests. More broadly, Hopp and Lovejoy (2012) show that in a typical imaging test unit, the time waiting on schedule is the single largest source of delay in getting images for most patients.

1.1.2 Literature Review

Our research continues the theme of expert services literature for which Dulleck and Kerschbamer (2006) provide an extensive review. Debo et al. (2008) model a monopolist expert who offers a service with unverifiable duration and hence has the incentive to delay the service. While embedding asymmetric information, their model does not address the differences in service quality. Paç and Veeraraghavan (2012) model the interaction between customers experiencing problems that can be either major or minor by nature, and an expert who may choose not to reveal the true nature of various problems in order to sell more extensive services. Debo and Veeraraghavan (2012), similar to us, assume that service time and service quality are positively correlated but they focus on analyzing strategic consumers' queue-joining behavior.

Our paper also draws from the literature on service management under congestion. Kostami and Rajagopalan (2009) analyze the intertemporal tradeoff between speed and quality in a general service setting. Tong and Rajagopalan (2012) study the pricing strategy for discretionary services when the service outcome is contractible and is directly driven by the service provider's service choice. Anand et al. (2011) study how a service provider resolves the tradeoff between service quality and speed when customers are strategic. They show that the customer intensity of the industry is a major determinant of the service provider's decision. Furthermore, they analyze the competition among multiple service providers and show that higher prices and service quality can result from more providers. Our paper differs from Anand et al. (2011) in that we consider insurance coverage in the healthcare market and emphasize the profound impact of insurance structure on the service usage under various service environments. Furthermore, we compare the actual and the socially efficient service consumption levels. Wang et al. (2010) develop a multi-server queueing model where a diagnostic service center that advises patients over phone. The service manager needs to strike a balance between accuracy of advice, callers' waiting time and staffing costs. Our paper addresses a similar tradeoff but focuses rather on the economic side of physicians' test-ordering behavior.

The supplier-induced demand (SID) literature contends that doctors as service providers, can directly influence patients' service usage. Patients seek advice from doctors largely because they cannot reach informed medical decisions on their own. While early SID models often view patients as perfectly informed but passive consumers, later studies treat patients as Bayesian decision-makers whose information-acquisition mechanism affects physicians' behavior. Our paper differs from the SID literature in three ways. First, SID models generally assume that physicians can observe patients' private information at no cost. Second, while waiting time limits the patients' access to healthcare, SID models treat it as a mechanism to control utilization and hence reduce the cost of *ex post* moral hazard (Gravelle and Siciliani 2008). As it is, the waiting time is the healthcare provider's unilateral decision rather than an outcome of physician-patient interaction. Third, the

SID literature typically assumes a fee-for-service payment model. For example, Sorensen and Gyrtten (1999) work on the premise that only contract physicians in Norway, whose incomes derive exclusively from patient visits or laboratory tests, have the incentive to induce demand. Our paper, by addressing the tradeoff among cost, access, and quality, justifies incentives to overtest even when more services do not imply additional revenue.

The paper proceeds in §1.2 with a model of strategic encounters between one physician and multiple patients, which enables us to characterize the market equilibrium and the social optimum. Section 1.3 analyzes the effect of insurance structure, reimbursement ceiling, and misdiagnosis concerns. Section 2.8 considers two extensions. The paper concludes in §1.5 with a summary of key operational and policy implications. All the technical proofs are relegated to the appendix.

1.2 Model

In this section, we model the interaction between a physician and a group of patients with exogenous demand under perfect information. We start by analyzing the relationship between test ordering and service quality. Then we model the tradeoffs faced by the patients and the physician. Finally, we characterize the market equilibrium and the social optimum, which gives the condition of overtesting.

1.2.1 Test Ordering and Service Quality

We capture the physician's test-ordering decision by the service rate μ that measures the speed of the overall diagnosis. This is based on our observation that a higher service rate results from demanding fewer tests, and vice versa, and allows us to capture the physician's problem under a strategic queueing theory framework.

In reality, the service rate of the system, which decreases in the number of tests is chosen from a discrete set.[1] For tractability, we assume that the service rate is a continuous variable. The service quality or the diagnostic certainty is determined by the consultation

[1] In the case of UPMC Eye Center, each ordered test is pre-assigned a slot with a fixed length, and the number of tests being ordered for a patient directly influences the service rate of conducting imaging tests.

session, as well as the physician's assessment of diagnostic results. The service quality given average service rate μ is defined as

$$Q(\mu) := Q_c + \alpha(\mu_c - \mu), \tag{1.1}$$

where Q_c denotes the baseline service quality; μ_c refers to the baseline service rate, that is, $Q(\mu_c) = Q_c$; α describes the rate in which the service quality improves when the service rate decreases.[2] It follows from (1.1) that $Q(\mu)$ decreases in μ, meaning that a slower service rate leads to higher service quality. This model is aligned with an elective outpatient setting such as an ophthalmology clinic, where additional imaging tests do not lead to the phenomenon of "over-diagnosis"; rather, more imaging tests "just increase the resolution of images so [physicians] can see them better" (Rheinbolt and Noecker 2009).

1.2.2 Patient Utility

Patients' utility from the service depends on 1) service quality, 2) waiting time, and 3) *out-of-pocket* payment. Patients are covered by indemnity insurance and pay less than the nominal service charge. There are several key components in a patient's health insurance plan: the deductible is the accumulative out-of-pocket expense to trigger insurance coverage; the copayment is the fixed charge that the patient must pay out of pocket for each visit; the coinsurance rate is the percentage of service fee, after accounting for the copayment, that the patient must pay out-of-pocket. We ignore the deductible to avoid the difficulty of defining the service fee below the deductible (cf. Newhouse 1978). All patients are assumed to have the same insurance coverage with zero deductible, a copayment of π, and a coinsurance rate of β. We assume homogeneous patients except in §1.4.1. The premium is viewed as a sunk cost and ignored. Letting p denote the nominal service fee, the patient's out-of-pocket payment is hence $\pi + \beta(p - \pi)$, since we focus solely on the interesting case where $p \geq \pi$.

[2] We use this function for simplicity of representation. When $Q(\mu)$ is a general concave and non-monotonic function of μ, we can show that our major insights continue to hold.

Patients arrive at an exogenous rate Λ, which is referred to as the potential demand for the service. Upon observing the physician's chosen service rate μ and service fee p, patients make queue joining decisions by adopting the following mixed strategies: each patient joins the queue with probability $\rho(\mu,p)$, and balks and resorts to an outside option with probability $1-\rho(\mu,p)$. Each patient's reservation utility is assumed to be zero without loss of generality. The induced arrival rate can be denoted as a function of μ and p such that $\lambda(\mu,p)=\rho(\mu,p)\cdot\Lambda$. This setting is consistent with the growing literature on equilibrium behavior of customers and servers in queueing systems (Hassin and Haviv 2003).

The potential demand for the service is assumed to follow a Poisson process, a reasonable representation for arrival processes in healthcare applications (Green 2006); thus, the induced arrival process resulting from patients' joint randomized decisions also follows a Poisson process. For simplicity, we assume that service time is exponentially distributed; our major results carry over to a general service time distribution. Hence, the service setting corresponds to an $M/M/1$ queue. Consistent with *money price* models (e.g., Coffey 1983), we define each patient's waiting time $W(\mu,\lambda)$ as the sum of the queueing time and the service time.[3] The expected waiting time in the $M/M/1$ queue is given by $W(\mu,\lambda(\mu,p))=1/[\mu-\lambda(\mu,p)]^{-1}$. Let ω denote the patient's waiting cost per unit of time. In practice, ω can be estimated as the value of lost productivity while waiting in the service queue (Phelps and Newhouse 1973; Coffey 1983). The sum of out-of-pocket expense $\pi+\beta(p-\pi)$ and waiting cost $\omega W(\mu,\lambda(\mu,p))$ is referred to as the full price. Then, using the market clearing condition

$$Q(\mu)=\pi+\beta(p-\pi)+\omega W(\mu,\lambda(\mu,p))$$

that equates the service quality to its full price and substituting for W, we obtain the induced arrival rate

$$\lambda(\mu,p)=\mu-\omega\left[Q(\mu)-\pi-\beta(p-\pi)\right]^{-1}.$$

[3] All of our major insights remain unchanged if we define the waiting time as the queueing time only.

In reality, the physician classifies patients into a number of pre-test types such that within each pre-test type, each patient is ordered an almost identical set of tests. Although such a patient mix faced by the physician is best described by a multiple classes of arrivals in a queueing network, we choose to focus on modeling patients who are classified as the same type, which ultimately influences the set of imaging tests. This allows us to analytically characterize patients' strategic queue joining decision in response to the service parameters.

1.2.3 Physician Behavior

We treat the physician as a price setter such that "the physician is assumed to have some control over the price he can charge and still obtain business" (Pauly 1980); this assumption is supported by patients' free choice of physicians, meaning that "in any negotiation over price between a physician and an insurer physicians have substantial bargaining power" (Newhouse 2002). The physician charges a fixed service fee p. The physician then chooses the service rate μ and the service fee p to maximize the revenue rate $g(\mu, p) = p\lambda(\mu, p)$. In addition, we assume $Q_c < \alpha\mu_c + (1-\beta)\pi$ to rule out the trivial case $\mu^* \geq \mu_c$. The assumption requires that the baseline service quality Q_c is lower than the sum of (1) $\alpha\mu_c = \lim_{\mu\to 0} Q(\mu) - Q_c$, the unattainable maximum service quality improvement, and (2) $(1-\beta)\pi$, each patient's copayment net of β, which is covered by the insurance. We characterize the equilibrium below.

Proposition 1.1 *There exists a unique market equilibrium where*

i) the physician chooses the service rate $\mu^* = [Q_c + \alpha\mu_c - (1-\beta)\pi]/(2\alpha)$, and the service fee $p^* = \frac{\alpha}{\beta}(\mu^* - \sqrt{\omega/\alpha})$.

ii) The induced arrival rate is $\lambda^* = \mu^* - \sqrt{\omega/\alpha}$.

iii) The average waiting time is $W^* = \sqrt{\alpha/\omega}$.

It is important to notice that the waiting time does not depend on the insurance structure. Since the waiting time $W(\mu, \lambda) = \left(\mu - \lambda\right)^{-1}$ spent per patient in the system depends only on the "surplus" service level $\mu^* - \lambda^*$, the optimal solution balances the cost α of increasing the service rate (i.e., the reduced diagnostic quality) to the reduction ω in patient's waiting costs.

1.2.4 Social Optimum and Overtesting Condition

The benchmark against which we measure overtesting is the social optimum that involves a social planner who determines the admission policy and the service rate to maximize the social welfare. Each physician-patient interaction generates a social surplus that is equal to the service quality, less patients' disutility from waiting. The expected social welfare rate is formulated as follows:

$$U(\mu,\lambda) = \lambda \cdot \left\{Q(\mu) - \omega W(\mu,\lambda)\right\}.$$

The following proposition gives the socially efficient service rate and arrival rate, denoted by μ^S and λ^S, respectively.

Proposition 1.2 *In the social optimum,*

i) the optimal service rate is $\mu^S = (Q_c + \alpha\mu_c)/(2\alpha)$;

ii) the optimal arrival rate is $\lambda^S = (Q_c + \alpha\mu_c)/(2\alpha) - \sqrt{\omega/\alpha}$;

iii) the expected waiting time is $W^S = \sqrt{\alpha/\omega}$.

Next, we compare the market equilibrium with the social optimum.

Corollary 1.1 *i) The physician orders no fewer tests in the market equilibrium than in the social optimum, that is,* $\mu^* \leq \mu^S$.

ii) The arrival rate is always greater in the market equilibrium, that is, $\lambda^* \leq \lambda^S$.

iii) The average waiting time is the same under the social optimum and in the market equilibrium, that is, $W^S = W^* = \sqrt{\alpha/\omega}$.

In the market equilibrium, the physician overtests due to the price distortions introduced by insurance coverage. This result is aligned with Feldstein's (1973) empirical finding that raising the coinsurance rate increases social welfare. In fact, when $\pi = 0$ and $\beta = 1$, patients are responsible for the entire payment, and the physician sets the service rate at the socially efficient level. Empirically, in Canada's health system, it was documented that provinces spending more on health care per person had neither shorter (nor longer) total waiting times than those spending less. In addition, those provinces spending more had lower rates of procedures and major surgeries (Zelder 2000), and increased spending was actually correlated with increases in waiting times unless those increases in spending were targeted to physicians (Esmail et al. 2003).

1.3 Analysis

This section analyzes the effects of insurance structure, reimbursement ceiling, and misdiagnosis concerns. We also briefly describe a robustness check of our key results.

1.3.1 Insurance Structure

We first examine the effect of the copayment and coinsurance components of the insurance plan on the physicians' test-ordering behavior. Proposition 1.1 implies the following result:

Corollary 1.2 *i) The physician's optimal service rate μ^* decreases in the copayment π and increases in the coinsurance rate β.*

ii) The physician's optimal service fee p^* decreases in both the copayment π and the coinsurance rate β.

The extant literature often suggests that increasing the patients' out-of-pocket expenses leads to decreased consumption of medical resources. The above corollary, by contrast, reveals that the copayment and the coinsurance rate can drive the consumption of imaging tests in *opposite* ways. In particular, the number of tests increases in the copayment π but decreases in the coinsurance rate β. To understand why, we examine each patient's out-of-pocket expense $\pi + \beta(p^* - \pi) = [Q_c + \alpha\mu_c + (1-\beta)\pi]/2 - \sqrt{\alpha\omega}$, which increases in π but decreases in β. As the copayment goes up, the physician needs to cut the service fee to ease the patients' monetary burden. Nevertheless, each patient's out-of-pocket expense still goes up because cutting the service fee by one dollar only reduces each patient's out-of-pocket expenses by $\beta < 1$ dollar, necessitating more tests to match the patients' increased monetary burden. With a higher coinsurance rate, however, the physician will charge a lower service fee, which leads to a reduced out-of-pocket expense for each patient, and justifies fewer tests ordered by the physician.

To the best of our knowledge, this is the first analytical finding about the impact of per-visit copayment on physicians' test-ordering behavior. There exist supporting empirical evidences for the result. Under an outpatient setting, Jung (1998) shows that increasing the per-visit copayment significantly reduces the number of office visits but increases the intensity of medical resource consumption for each visit.

Next, we examine the effect of the insurance structure on the social welfare gap between the market equilibrium and the social optimum. The social welfare gap, written as a function of β and π, is $\Delta U(\pi,\beta) = U(\mu^S,\lambda^S) - U(\mu^*,\lambda^*) = \pi^2(1-\beta)^2/(4\alpha)$, and its second-order derivatives in terms of β and π are $\partial^2\Delta U/\partial\beta^2 = \pi^2/(2\alpha) \geq 0$, and $\partial^2\Delta U/\partial\pi^2 = (1-\beta)^2/(2\alpha) \geq 0$, respectively. Hence we have the following corollary:

Corollary 1.3 *The social welfare gap is convex decreasing in the coinsurance rate β, and convex increasing in the copayment π.*

As the copayment increases, the physician tends to order more tests for each patient but the induced arrival rate decreases. Combining the decreased arrival rate with higher number of tests per patient visit, we observe that more resources are consumed by fewer individuals at any given time, widening the social welfare gap at a faster pace. This phenomenon explains why the social welfare gap is convex increasing in the copayment π. As the coinsurance rate increases, the physician's test-ordering pattern and the equilibrium arrival rate converge to the socially efficient one.

1.3.2 Reimbursement Ceiling

Given that insurance coverage distorts the demand curve for imaging services, one natural proposal is to introduce a reimbursement ceiling which sets the maximum reimbursable amount for each service session. This essentially restricts the maximum service fee charged by the physician, which we denote by $p_{\max}$. Defining the maximum out-of-pocket expense by $q_{\max} = \pi + \beta(p_{\max} - \pi)$, the equilibrium is characterized in the proposition that follows.

Proposition 1.3 *Depending on the size of $p_{\max}$, two possible equilibrium outcomes can arise:*

i) If $p_{\max} > [Q_c + \alpha\mu_c - 2\sqrt{\omega\alpha} - (1-\beta)\pi]/(2\beta)$, then the equilibrium is characterized by Proposition 1.1.

ii) If $p_{\max} \leq [Q_c + \alpha\mu_c - 2\sqrt{\omega\alpha} - (1-\beta)\pi]/(2\beta)$,

a) the physician chooses the service fee $p^* = p_{\max}$ and the service rate $\mu^* = (Q_c + \alpha\mu_c - q_{\max})/\alpha - \sqrt{\omega/\alpha}$;

b) the induced arrival rate is $\lambda^* = \mu^* - \sqrt{\omega/\alpha}$;

c) the average waiting time is $W^* = \sqrt{\alpha/\omega}$.

When $p_{\max}$ is low, it becomes restricting such that physician would choose a service fee that is exactly the same as $p_{\max}$. This essentially corresponds to the scenario where the physician's service fee is capped by the insurance firm's reimbursement policy.

The following corollary illustrates how the presence of a reimbursement ceiling affects the physicians' test-ordering behavior.

Corollary 1.4 *i) The physician's optimal service rate μ^* decreases in the copayment π.*

ii) The physician's optimal service rate μ^* increases in the coinsurance rate β if and only if the reimbursement ceiling exceeds $p_{\max} > [Q_c + \alpha\mu_c - 2\sqrt{\omega\alpha} - (1-\beta)\pi]/(2\beta)$.

The intuitions behind Corollary 1.4 are three-fold. First, *ceteris paribus*, when the copayment increases, the physician compensates patients' utility loss by ordering more tests. Second, when the reimbursement ceiling $p_{\max}$ is high enough, greater insurance coverage encourages overtesting, as patients are less sensitive to the service fee. That is, the physician responds to a decrease in the coinsurance rate β by ordering more tests. Third, when the insurance firm sets a low reimbursement ceiling $p_{\max}$, the physician will set the service fee at exactly $p_{\max}$. A lower coinsurance rate β, similar to a lower copayment π, reduces patients' fixed out-of-pocket payment, and the physician can order fewer tests without sacrificing patients' net surplus. Therefore, under a low reimbursement ceiling, increasing patients' cost-sharing can lead to a higher testing level.

We briefly discuss the social welfare gap based on Corollary 1.4. As in the baseline model, the social welfare gap is convex increasing in the copayment π since both the service rate μ^* and the equilibrium arrival rate λ^* decrease in π. The social welfare gap is convex decreasing in β when the reimbursement ceiling $p_{\max}$ is high, as in the baseline model. With a low reimbursement ceiling, however, both the arrival rate and the service rate decrease in β, meaning that the social welfare gap is convex increasing in β.

The following corollary compares the market equilibrium with the social optimum.

Corollary 1.5 *i) If the reimbursement ceiling $p_{\max} > [Q_c + \alpha\mu_c - 2\sqrt{\omega\alpha} - (1-\beta)\pi]/(2\beta)$, then the physician always orders more tests than the socially efficient level, that is, $\mu^* < \mu^S$.*

ii) If the reimbursement ceiling $p_{\max} \leq [Q_c + \alpha\mu_c - 2\sqrt{\omega\alpha} - (1-\beta)\pi]/(2\beta)$, then the physician can order more or fewer tests than the socially efficient level, that is, both $\mu^* \leq \mu^S$ and $\mu^* > \mu^S$ are possible.

The above corollary provides the condition under which overtesting occurs. When the reimbursement ceiling is sufficiently high, the physician always overtests. With a low reimbursement ceiling, however, the physician can either overtest or undertest, depending on whether each patient's out-of-pocket expense $q_{\max}$ is over $(Q_c + \alpha\mu_c - 2\sqrt{\alpha\omega})/2$. This is because a higher net payment is compensated by more tests, and vice versa.

Corollary 1.5 also helps uncover the puzzle that motivates our research. Recall from §1.1 that, overtesting occurs even under the exogenous pricing scenario, that is, when the physician receives the same income per patient visit regardless of the number of tests ordered. Consider a setting in which the physician's compensation per patient visit is fixed at $\bar{p}$. The service rate becomes the physician's sole decision. This problem is equivalent to the case where the reimbursement ceiling is set low enough, and the physician always sets the service fee at the maximum possible amount. In equilibrium, the physician chooses a service rate of $\mu^* = [Q_c + \alpha\mu_c - \pi - \beta(\bar{p} - \pi)]/\alpha - \sqrt{\omega/\alpha}$, which can be either higher or lower than the socially efficient service rate μ^S. In other words, overtesting is still possible even under an exogenous pricing system.

1.3.3 Misdiagnosis Concerns

There are scenarios where the physician bears the risk of misdiagnosis. The physician is subject to a penalty if there exists substantial proof that a patient's condition has worsened because the physician failed to interpret the testing results accurately and therefore did not act on time. In some other cases, an inadequate number of tests can indicate that a normal patient is abnormal, exposing patients to unnecessary treatments. Prior medical literature validates the significance of misdiagnosis concerns in their scope and impacts. Studdert et al. (2006) find that 37% of malpractice claims do not involve any *real* medical errors but account for 13–16% of the system's total costs. In a study to reveal physicians' perceived risk of misdiagnosis, Carrier et al. (2010) confirm high malpractice concerns among physicians at all levels even when malpractice risks are sufficiently low by objective measures. They also find that such concerns are not eased by common tort reforms. Baicker et al. (2007) show that increased malpractice risk drives higher consumption levels of healthcare services, especially in discretionary services.

We model the physician's misdiagnosis concerns as a simple misdiagnosis cost function of the service rate: $\theta(\mu) := d \cdot \mu$, where d is a constant denoting the marginal misdiagnosis cost. The misdiagnosis cost increases in μ, to align with the observation that fewer tests cause the physician to be more concerned about the possibility of reaching an inaccurate diagnosis. When μ is very small, indicating that the physician orders a sufficiently large number of tests, the misdiagnosis cost approaches zero.

The physician's decision consists of choosing the service rate $\mu \in (0, \mu_c)$ and the service fee p to maximize the utility rate $g_m(\mu, p) = [p - \theta(\mu)] \cdot \lambda(\mu, p)$. We characterize the equilibrium in the following proposition:

Proposition 1.4 *In the case with misdiagnosis concerns,*

i) the physician chooses the service rate $\mu_m^* = [Q_c + \alpha\mu_c - (1-\beta)\pi]/[2(\alpha+\beta d)]$*, and the service fee* $p_m^* = (\alpha + 2\beta d)[\mu_m^* - \sqrt{\omega/\left(\alpha+\beta d\right)}]/\beta$*;*

ii) the induced arrival rate is $\lambda_m^* = \mu_m^* - \sqrt{\omega/\left(\alpha+\beta d\right)}$*;*

iii) the average waiting time is $W_m^* = \sqrt{\left(\alpha+\beta d\right)/\omega}$*.*

The corollary below follows from Proposition 1.4.

Corollary 1.6 *i) With misdiagnosis concerns, the physician's optimal service rate* μ_m^* *decreases in the copayment* π*.*

ii) If $d < \alpha[(Q_c + \alpha\mu_c)/\pi - 1]^{-1}$*, then the physician's optimal service rate* μ_m^* *increases in the coinsurance rate* β*; otherwise, the physician's optimal service rate* μ_m^* *decreases in the coinsurance rate* β*.*

An increase in the fixed per visit charge increases the expectation for service quality and so justifies more tests. An increase in the coinsurance rate β, however, can lead to either an increase or a reduction in the optimal service rate μ^* depending on d. When d is low, similar to the case without misdiagnosis concerns, an increase in the coinsurance rate leads to a lower service fee and lower service quality and hence increase in the service rate. When d is high, due to the high misdiagnosis cost, the physician no longer finds it optimal to lower the service quality due to a high misdiagnosis concern. Rather, it is optimal to compensate patients' higher expenses by increasing the service quality (i.e., decreasing the service rate).

Next, we derive the condition under which the physician would overtest. The social planner aims to maximize the social welfare rate that can be represented as $U_m(\mu,\lambda) = \lambda\cdot\left\{Q(\mu) - \theta(\mu) - \omega W(\mu,\lambda)\right\}$. The next proposition characterizes the social optimum.

Proposition 1.5 *With misdiagnosis concerns, in the social optimum,*

i) the optimal service rate is $\mu_m^S = \frac{Q_c+\alpha\mu_c}{2(\alpha+d)}$*;*

ii) the optimal arrival rate is $\lambda_m^S = \mu^S - \sqrt{\omega/(\alpha+d)}$*;*

iii) the expected waiting time is $W_m^S = \sqrt{(\alpha + d)/\omega}$.

The following corollary is immediate from Propositions 1.4–1.5.

Corollary 1.7 *If the copayment* π *is higher than* $(Q_c + \alpha\mu_c)/(1 + \alpha/d)$*, then the physician orders more tests than the socially efficient level, that is,* $\mu_m^* < \mu_m^S$*; otherwise, the physician orders fewer tests than the socially efficient level.*

Corollary 1.7 is rather counterintuitive: when physicians are concerned by potential inaccurate medical judgment, they can order either more or fewer tests than the socially efficient level (the latter case is referred to as "undertesting"). It is especially surprising in view of Corollary 1.1 which states that the physician always overtests in the absence of misdiagnosis concerns. To understand this result, we need to examine Proposition 1.5, which shows that misdiagnosis concerns increase the socially efficient consumption level. The insurance coverage, on the other hand, enables patients to pay less than the actual service fee. Specifically, when the copayment is lower than $(Q_c + \alpha\mu_c)/(1 + \alpha/d)$, the physician can satisfy patients by ordering fewer tests than the socially efficient level. When the copayment exceeds $(Q_c + \alpha\mu_c)/(1 + \alpha/d)$, the physician's efforts to induce demand are supplemented by the insurance coverage. Furthermore, given Q_c and μ_c, the threshold decreases in the ratio of α and d. Consider the special case where the physician's misdiagnosis concern is sufficiently low (d is small): the threshold is then close to zero, meaning that the physician invariably overtests, which is consistent with Corollary 1.1.

Corollary 1.8 *The average waiting time in the social optimum is longer than in the market equilibrium, that is,* $W_m^S > W_m^*$.

Corollary 1.8 may initially seem surprising in that, even when the physician orders more tests than the socially efficient level, patients still experience a shorter expected waiting time. The underlying intuition is as follows. We first recognize that one way to implement the social optimum is to charge each patient a service fee coinciding with the patient's externality by joining the queue

$$p_m^S = Q(\mu_m^S) - \omega W_m^S = (\alpha + 2d)(Q_c + \alpha\mu_c)/[2(\alpha + d)] - \sqrt{\omega(\alpha + d)}. \tag{1.2}$$

Under the market equilibrium, however, each patient's out-of-pocket expense is

$$\pi + \beta(p_m^* - \pi) = \frac{(\alpha + 2\beta d)(Q_c + \alpha\mu_c) + \alpha\pi(1-\beta)}{2(\alpha + \beta d)} - \sqrt{\omega(\alpha + \beta d)}. \tag{1.3}$$

Recall from Corollary 1.7 that, when $\pi > (Q_c + \alpha\mu_c)/(1 + \alpha/d)$, the physician overtests. In the meanwhile, comparing (1.2) and (1.3) gives that $\pi + \beta(p_m^* - \pi) > p_m^S$, meaning that each patient is subject to a high out-of-pocket expense, which essentially induces a low arrival rate. Consequently, the gap between the induced arrival rate and the service rate is higher than under the social optimum, leading to a lower expected waiting time. This phenomenon has been observed empirically in Hong Kong's healthcare system where the average waiting times across public and private hospitals are significantly different: the average waiting time for a physician in public hospitals is 74.7 days, whereas it is 24.3 days for a private physician (Harvard Team 1999).

1.3.4 Robustness Check

Our model makes a few simplifying assumptions to sharpen the managerial insights. To explore the robustness of our main results, we conduct a thorough investigation into various extensions. These additional extensions are available upon request.

- **A general, non-monotonic service quality function.** Our model assumes the service quality $Q(\mu)$ to be an affine function of μ for simplicity of analysis. When we extend $Q(\mu)$ to a general function form and allow it to be non-monotonic, that is, excessive testing leads to decrease in the quality of service (Gupta 2012), we can show that all of our major insights hold in the full-information case.

- **Distribution of service time.** Throughout the paper we assume that the service time is exponentially distributed. We show that all of our results pertaining to the physician's test-ordering behavior remain valid when the service time follows a general distribution. Nevertheless, different coefficients of variation are reflected in the service fee.

- **Definition of waiting time.** We have thus far used the total time in system as the definition of the waiting time since the benefits from the service have been captured by the function $Q(\mu)$. In the case in which the waiting time is defined as the time in the queue only, we can show that our major results remain unchanged.

- **Misdiagnosis cost function.** Our model assumes an affine misdiagnosis cost function. When this assumption is relaxed to a general function that is convex increasing in μ, our insights regarding misdiagnosis remains unchanged.
- **Diagnosis time**. Physicians need to spend time reading and analyzing test results, which in essence shortens their effective service time. The tradeoff here is between maximizing service throughput and improving service quality. We can show that our major findings carry over when considering the physician's diagnosis time.
- **Follow-up visits.** We assume in our model that at any time every patient with medical needs adopts the same randomized queue-joining strategy; in consequence, a patient can make multiple visits, but different visits are deemed independent of each other. In practice, however, a proportion of patients making initial visits are advised by the physician to make follow-up visits. We can show that although the physician exhibits different test-ordering patterns for new and returning patients, our major findings remain directionally valid.
- **Effort-averse physician.** The labor costs involved in the analysis of test of results are ignored in our model. When the physician is effort-averse, we find that the effort costs are reflected in the service fee rather than the physician's test-ordering pattern.

1.4 Extensions

In this section, we consider two extensions on patient heterogeneity and information asymmetry, respectively.

1.4.1 Patient Heterogeneity

The baseline model with homogeneous patients provides a benchmark for understanding physicians' test-ordering behavior. Now we consider the case where patients can have different insurance coverage and valuations for time.

There are two patient groups, those who have "good" insurance plans are denoted by g and those who have "bad" insurance plans are denoted by b. Type i patients' health plans are specified by a copayment π_i and a coinsurance rate β_i, $i = g, b$. We assume that $\pi_g < \pi_b$ and $\beta_g < \beta_b$ so that patients with good insurance have lower out-of-pocket expenses given the same nominal service fee. The potential arrival rate of group i is Λ_i, which is sufficiently large such that full coverage is not possible. Within each patient group, patients differ in their sensitivity to delay, that is, their waiting costs. Furthermore, we assume that a type $i \in \{g, b\}$ patient has a waiting cost, denoted by ω_i that is uniformly distributed in $[\underline{\omega}_i, \bar{\omega}_i]$. We focus our attention on the case where $\bar{\omega}_g > \bar{\omega}_b$ and $\underline{\omega}_g > \underline{\omega}_b$ such that patients with good health plans have higher time prices.

We assume that the physician cannot discriminate among patients by adopting various appointment intervals (and hence service rates) or service fees based on patients' valuation for money (dictated by insurance coverage) or valuation for time.[4] This reflects the case that the physician does not possess each patient's specific insurance information and thus chooses to base her test-ordering decisions on the profile of the "average" patient who seeks service. When the physician chooses the service rate μ and the service fee p, there exist critical levels ω_i^* such that only type i patients with $\omega_i \leq \omega_i^*$ for $i = g, b$ join the queue; the other patients opt out of the queue and seek for outside options. Assuming zero reservation utility for each patient type, ω_g^* and ω_b^* are determined by solving the following two equations in which λ_g and λ_b correspond to the equilibrium arrival rates of the two types:

$$Q(\mu) - P_i(p) - \frac{\omega_i^*}{\mu - \lambda_g - \lambda_b} = 0 \text{ for } i = g, b, \tag{1.4}$$

$$\lambda_i = \Lambda_i \left(\omega_i^* - \underline{\omega}_i \right) / \Delta\omega_i \text{ for } i = g, b, \tag{1.5}$$

[4] Indeed, media has exposed cases where different insurance types (private versus Medicare) were quoted different times to access, which lead to embarrassment and changes, cf. Grady (2011).

where $P_i(p) = \pi_i + \beta_i(p - \pi_i)$ and $\Delta\omega_i = \bar{\omega}_i - \underline{\omega}_i$ for $i, j \in \{g, b\}$, $j \neq i$. Jointly solving (1.4)–(1.5) gives the equilibrium arrival rates when the physician chooses μ and p:

$$\lambda_i(\mu, p) = \frac{\rho_j\left[\mu P_j(p) - \mu Q(\mu) + \underline{\omega}_i\right] + Q(\mu)\Delta\omega_i + P_j(p)\underline{\omega}_j - P_i(p)\underline{\omega}_i}{\sum_{k=g,b}\rho_k[P_k(p) - Q(\mu)] - \rho_i\rho_j},$$

where $\rho_i = \Delta\omega_i/\Lambda_i$ for $i \in \{g, b\}$. The physician's objective function is then given by $g(\mu, p) = p\left[\lambda_b(\mu, p) + \lambda_g(\mu, p)\right]$.

To facilitate our analysis, we define the quantity $\mu_i^* = [Q_c + \alpha\mu_c - (1 - \beta_i)\pi_i]/(2\alpha)$ for $i = g, b$, which corresponds to the optimal service rate when there exist only type i patients with homogeneous waiting costs, cf. Proposition 1.1.

Using a similar procedure to the one used in the proof of Proposition 1.1, we obtain the optimal service rate for the heterogeneous-patients system as follows:

$$\mu^* = \frac{\rho_g\mu_g^* + \rho_b\mu_b^*}{\rho_g + \rho_b} - \sqrt{\left(\frac{\rho_g\mu_g^* + \rho_b\mu_b^*}{\rho_g + \rho_b}\right)^2 - \frac{\rho_g\underline{\omega}_b + \rho_b\underline{\omega}_g}{\alpha\left(\rho_g + \rho_b\right)}}.$$

One would expect that μ^* is between μ_b^* and μ_g^* under a mix of both types of patients. The following proposition, however, suggests that heterogeneity in patient type always induces the physician to order more tests than the homogeneous case.

Proposition 1.6 $\mu^* < \min\{\mu_b^*, \mu_g^*\}$.

Next, we analyze the social optimum under patient heterogeneity. Since both Λ_b and Λ_g are sufficiently large, the optimal admission control policy is dictated by a parameter $\hat{\omega}$ such that only patients with waiting costs lower than $\hat{\omega}$ are admitted into the queue; insurance coverage no longer plays a role.

We use μ_h^S to denote the socially efficient service rate given the heterogeneity. The socially efficient service rate μ^S under homogeneity, recall from Proposition 2, is independent of the waiting cost. The following Proposition summarizes the effect of patient heterogeneity on the socially efficient service rate:

Proposition 1.7 *The socially efficient service rate in the heterogeneous system is always lower than in the homogeneous system, i.e., $\mu_h^S < \mu^S$.*

We now compare the market equilibrium with the social optimum. From the optimality conditions for the market equilibrium, we derive that $\omega_b^* < \omega_g^*$, that is, the marginal type g patient suffers more from delay than the marginal type b patient. This implies that $\lambda_g/\lambda_b > \Lambda_g/\Lambda_b$ when translated into the equilibrium arrival rates. That is, the physician distorts the fraction of type g patients that she sees compared to the population average, leading to the phenomenon of *selection* behavior (cf. Newhouse 2002). Moreover, notice that when $\Delta\omega_g \geq \Delta\omega_b$, that is, the type g patient has more variable waiting cost, we have $\omega_g^* > \omega_b^* + \left(\underline{\omega}_g - \underline{\omega}_b\right)$. In other words, not only does the marginal type g patient have a higher delay cost rate than the marginal type b patient, but also the physician finds it optimal to see a disproportionate fraction of type g patients compared to the population average since $\lambda_g/\lambda_b > \Lambda_g/\Lambda_b$. In contrast, in the social optimum, the waiting cost rate of the marginal patients for both types is equal to the common threshold $\hat{\omega}$. Therefore, one might expect the average waiting time in the market equilibrium to be lower than in the social optimum since more can be charged to type g patients due to the drop in waiting times: type g patients are not only more delay-sensitive, they can also absorb a higher price increase because of their better insurance coverage. While the waiting cost helps increase the service rate under both equilibria, this decrease in the waiting cost is experienced mainly by the type g patients in the market equilibrium through the increase in fees collected per unit time. This is consistent with the general view from welfare economics that market equilibrium leads to under-utilization of a system than is optimal from the social planner's point of view.

We close this section by briefly discussing the impact of patient heterogeneity on the social welfare gap. The social planner's objective is to maximize the social welfare and therefore does not place any weight on the individual's insurance coverage when determining the admission policy and the service rate. However, the physician has the incentive to choose the service rate and the service fee to cherry-pick a less price-sensitive mix of patients. The difference is clearly reflected in the fact that there exist two cut-off waiting costs, namely, $\omega_i^*, i = g, b$, under the market equilibrium, but only one $\hat{\omega}$ in the social optimum. The higher the difference between ω_g^* and ω_b^*, the wider is the social welfare gap between the market equilibrium and the social optimum.

1.4.2 Physician Type Uncertainty

In this section, we model the encounters between patients and the physician under initial uncertainty of the physician's skill level. In contrast to the preceding sections, the physician's skill level, referred to as "type," is unobservable to patients. The physician's type is denoted by $s \in \{h, l\}$, and a type s physician's skill level is α_s. We assume that $\alpha_h > \alpha_l$, indicating that, given the amount of tests, the service provided by a type h physician yields higher diagnostic certainty. Unaware of the physician's type, patients are provided with access to the physician's pricing information before choosing a physician.

The interaction between the physician and patients lasts for two service periods. Figure 1.2 provides a time line depicting the sequence of events. At $t = -1$, the physician discovers her own type s, which can be either h (with prior probability ψ_0) or l (with prior probability $(1 - \psi_0)$). Patients are perfectly informed of the distribution of the physician's type but not its realization. At $t = 0$, the physician sets her service fee p_s, which remains unchanged thereafter. After observing the posted service fee p_s, patients form their posterior beliefs $\Psi(p_s)$ that the physician is of type h. Anticipating patients' beliefs about her type, type $s \in \{l, h\}$ physician chooses the service rate μ_{s1} in the first period. The equilibrium arrival rate $\lambda_1(p_s)$ during the first service period, is affected by both patients' posterior beliefs and the physician's chosen service rate, and can be solved from the following equation:

$$Q_h(\mu_{s1})\Psi(p_s) + Q_l(\mu_{s1})(1 - \Psi(p_s)) - \omega W(\mu_{s1}, \lambda_1(p_s)) - \pi - \beta(p_s - \pi)^+ = 0, s = l, h,$$

where $Q_s(\mu) = Q_c + \alpha_s(\mu_c - \mu)$ for $s \in \{h, l\}$ represents the service quality given the service rate μ when the physician type is h and l, respectively. At $t = 2$, the physician type is revealed to patients through physician's first period choice of service rate μ_{s1} and the service quality delivered. Therefore, in the second period, the queue-joining decisions faced by patients, as well the service rate decision faced by the physician, are similar to those in the baseline model.

We make the following two assumptions to facilitate our analysis:

Assumption 1.1 $\mu_c \geq \sqrt{\omega/\alpha_l}$.

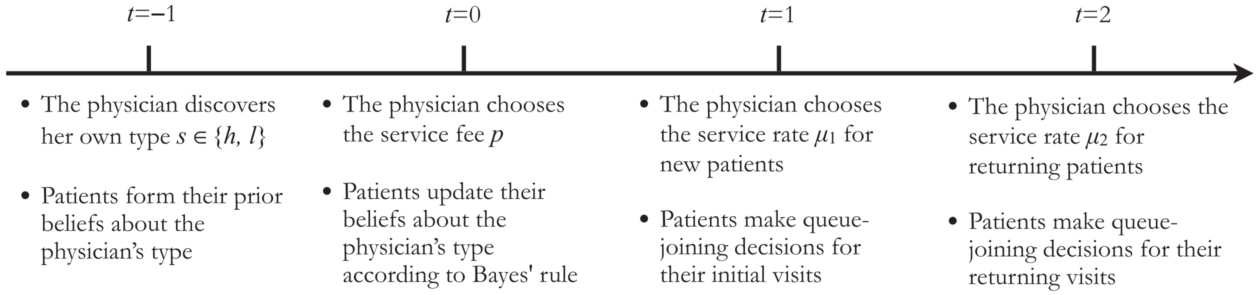

t=−1
t=0
t=1
t=2
The physician discovers her own type s ∈ {h, l}
Patients form their prior beliefs about the physician's type
The physician chooses the service fee p
Patients update their beliefs about the physician's type according to Bayes' rule
The physician chooses the service rate μ1 for new patients
Patients make queue-joining decisions for their initial visits
The physician chooses the service rate μ2 for returning patients
Patients make queue-joining decisions for their returning visits

Figure 1.2: Time line of the model with physician type uncertainty

Assumption 1.2 $Q_c + \alpha_s \mu_c - 2\sqrt{\alpha_s \omega} \geq (1+\beta)\pi$.

Assumption 1 rules out the uninteresting cases of $p_s^* < \pi$ and Assumption 2 ensures that $1/\mu_c \leq \sqrt{\alpha_s/\omega}$ for $s = \{l, h\}$, i.e., the average waiting time in market equilibrium for both types of physicians is at least as large as that under the base service rate of μ_c.

Full information. Under full information, the patient would be perfectly informed of the physician's type and the equilibrium would coincide with that in the baseline model. We denote by (μ_s^*, p_s^*) the type s physician's optimal decisions under full information for $s \in \{h, l\}$. Then, Proposition 1 implies that both types of physicians overtest in the full-information equilibrium. In the next corollary, we compare revenue rates of the two types of physicians under the full-information equilibrium.

Corollary 1.9 *If physician types were observable by patients, then the high type physician's expected revenue rate would always be higher than that of the low type.*

Corollary 1.9 suggests that, when the physician type is unobservable to patients, the type h physician prefers to be separated from the type l physician, while the type l physician prefers not to be separated from the type h physician. In other words, the type l physician has an incentive to *mimic* the type h physician.

Asymmetric information. We model the physician-patient interaction as a sequential game of incomplete information (see Debo et al. 2012 for another example of a signalling game in a queueing context). We restrict attention to pure strategy perfect Bayesian equilibria satisfying the Intuitive Criterion of Cho and Kreps (1987), consistent with most of the literature on signaling games. The intuitive criterion is an equilibrium refinement which restricts beliefs off the equilibrium path. In particular, it requires that the updating of beliefs should not assign positive probability to a player taking an action that is equilibrium dominated. Essentially, the intuitive criterion allows us to eliminate any perfect Bayesian equilibrium from which some type of physician would want to deviate even if she were not sure what exact belief the patients would have as long as she knows that the patients would not think she is a type who would find the deviation equilibrium dominated.

We next characterize the equilibrium for the signalling game and discuss its implication for the physician's test-ordering behavior. In the following proposition, we show that the Intuitive Criterion selects a unique outcome, in which the unobservable physician skill information is fully revealed to the patients through fees charged by the physicians. While in certain cases type *h* physician is able to achieve such a separation without sacrificing her full-information revenue rate, in other cases the type *h* physician needs to deviate from her full-information strategy.

Proposition 1.8 *The unique equilibrium of the signalling game under the Intuitive Criterion is the minimal-cost separating outcome characterized by the threshold skill difference* $\underline{\Delta\alpha}= \alpha_H - \alpha_L > 0$ *defined by*

$$\mu_c\left(\sqrt{\alpha_h/\alpha_l}-1\right)\Delta\alpha - 2\sqrt{\alpha_l\omega}\left(\sqrt{\alpha_h/\alpha_l}-1\right)^2 = Q_c + \alpha_l\mu_c - \pi\left(1-\beta\right) - 2\sqrt{\alpha_l\omega}.$$

a. ***Costless separating equilibrium.*** *If* $\Delta\alpha \geq \underline{\Delta\alpha}$*, in the unique separating equilibrium the type l physician charges* p_l^* *and the type h physician charges* p_h^**. The patients beliefs are:* $\Psi(p_l^*) = 1$ *and* $\Psi(p) = 0$ *otherwise.*

b. ***Costly separating equilibrium.*** *If* $\Delta\alpha < \underline{\Delta\alpha}$*, in the unique separating equilibrium the type l physician charges* p_l^* *and the type h physician charges* $p_h' > p_h^*$ *such that*

$$p_h' = \frac{\alpha_l p_h^* + \alpha_h p_l^*}{\alpha_l + \alpha_h} + \sqrt{\left(\frac{\alpha_l p_h^* + \alpha_h p_l^*}{\alpha_l + \alpha_h}\right)^2 - \left(\alpha_h p_l^*\right)^2}$$

the patient's beliefs are: $\Psi(p) = 1$ *if* $p = p_h'$*;* $\Psi(p) = 0$ *otherwise.*

Proposition 1.8 shows that in the signaling game between the physician and the patients, the only outcome is a separating equilibrium. Hence, the physicians' use of the service fees as signals of their skill is a self-selection mechanism that generates information "endogenously." The signal conveys the appropriate message to the patients, who use this mechanism to their advantage in order to select high-skilled physicians. Since both sides of the market are satisfied with the workings of the mechanism, a unique separating equilibrium emerges. Any possible pooling equilibrium is ruled out since the high type can always exploit the economic benefits of providing higher diagnostic certainty in order to separate from the low type while such a deviation would be dominated for the low type.

Furthermore, this separating equilibrium is the least-distortive one. Because a high-skill physician is more willing to trade off an increase in service fee and tests to induce an increase in revenue, it is possible to find a suitable service fee level that is worth charging if and only if the physician's relative skill level is high. Under the costless separating equilibrium, both types of physicians behave as if in the full-information equilibrium. When the physicians' skill level difference $\Delta\alpha$ is low enough, the type l physician has the incentive to mimic the type h physician. In this case, a separating equilibrium prevails if the type h physician manages to signal her type by deviating from the full-information equilibrium. The signal is said to be *costly* because the type h physician sacrifices a proportion of her revenue to deter the type l physician from mimicking. Under the costly separating equilibrium, the type h physician chooses a service fee higher than p_h^* to signal her type. In the meantime, the type h physician orders more tests than in the full information equilibrium to compensate for patients' utility loss. The signal is said to be *costly* because the type h physician sacrifices a proportion of her revenue to deter the type l physician from mimicking.

Three observations may now be made.

Observation 1.1 Price transparency might lead to higher service fees.

The opacity in pricing in health care services is a well-known phenomenon that separates the health care industry from markets for most goods and services. The pending

Transparency in All Health Care Pricing Act of 2010 will require all the health care providers (including hospitals, physicians, nurses, etc.) to post prices for various services (Kiviat 2010). Notwithstanding many intuitive benefits associated with price transparency, the Congressional Budget Office (2008), by citing empirical evidences from other industries, contends that increasing transparency in the healthcare market can result in higher prices. Proposition 1.8 indicates that, with price transparency and the increasingly popular consumer practice of "comparison shopping" in the healthcare sector (Miller 2010), a pooling equilibrium, in which both types of physicians choose a medium service level, can never sustain; price transparency encourages the type h physician to overtest, and prevents the type l physician from mimicking the type h physician. By comparison, under pricing opacity, a pooling equilibrium sustains. This gives an implication similar to the finding by the Congressional Budget Office (2008) albeit from a different angle: price transparency leads to higher prices and encourages the prescription of unnecessary tests.

Observation 1.2 *Without signalling opportunities, the low skill physician would always be better off whereas the high-skill physician would be better off if* the fraction ψ_0 *of type h physicians in the profession is high and the skill difference is low.*

Compared to the case in which signalling is not available, the low-skill physician is worse off. Without signalling, the physician would be treated as if she possessed the average skill and she would see a patient demand accordingly. Here instead, the low-skill physician earns only his full information revenue. Surprisingly, the high-skill physician may also be worse off. As no increase in service fee (and an accompanying increase in number of tests) is interpreted as evidence of low skill, the outcome without signalling is not available to her any longer. In a costly separating equilibrium, she earns less than the full information revenue. Without signalling opportunities, her revenue is as if she had average skill. While this tends to as full information benchmark for type h as ψ_0 tends to one, the separating equilibrium is independent of ψ_0. Therefore, if ψ_0 is large enough, the high-skill physician is worse off. as well.

Observation 1.3 Improved diagnostic technology can either exacerbate or alleviate the phenomenon of overtesting.

The medical community has divided views regarding whether improved technology will increase or reduce healthcare expenditure and the social welfare (Newhouse 1992). Newhouse (1992) cites technological change as a major escalator of health care expenditure. Cutler and McClellan (2001) examine five medical conditions (breast cancer, cataracts, depression, heart attacks, and low-birthweight infants), and argue that the benefits associated with improved technology outweigh increased costs.

We conduct numerical experiments to understand the impact of improved diagnostic technology. We maintain the type h physician's skill level and increase the type l physician's skill level gradually to reflect the notion that improved technology flattens out the skill level differences among physicians. The results, as shown in Figure 1.3, illustrate that technology advancements can either exacerbate or alleviate the phenomena of overtesting depending on the range of the skill level differences between the two types of physicians:

- Region I: the physicians' skill level difference is high (α_l is low). In this case, the costless separating equilibrium prevails, and the improvement in diagnostic technology has little impact on the physician's test-ordering behavior.
- Region II: the physicians' skill level difference is medium (α_l is medium). In this case, the type h physician uses a costly signal to separate from the type l physician. As technological advancements lead to less differentiation between physicians in the level of skill, even though patients achieve diminishing service quality gains by switching from the type l physician to the type h one, the type h physician has an even stronger incentive to overtest as a costly signaling effort. In other words, the improvement of diagnostic technology leads to more salient overtesting behavior.
- Region III: the physicians' skill level difference is low (α_l is comparable to α_h). The costly separating equilibrium continues to prevail, but an increased α_l makes it less rewarding for the type h physician to signal her type. As a consequence, the improvement in diagnostic technology leads to a lower incentive for overtesting.

We now briefly discuss the social welfare gap under asymmetric information about the physician type. Better insurance coverage, characterized by a low copayment π and a low coinsurance rate β, reduces patients' price-sensitivity and makes it more likely for the physician to adopt a costly signal. Consequently, in contrast to the baseline model, a higher

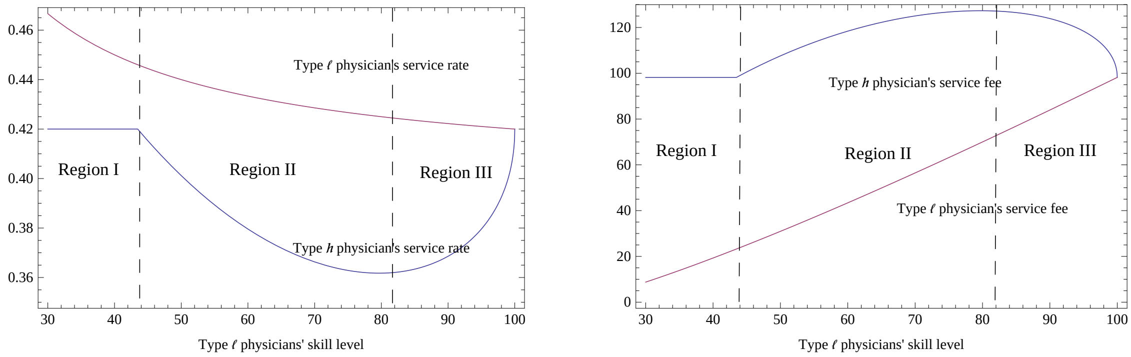


Figure 1.3: The impact of the different skill level differences among physicians on the physicians' service rates and fees.

Note: Parameters are: $\mu_c = 0.8, Q_c = 8, \omega = 5, \pi = 5, \beta = 0.2, \alpha_h = 100$.

copayment π can be social-welfare-improving as it can serve to undermine physicians' signaling efforts.

1.5 Concluding Remarks

To our best knowledge, this work is the first to analytically investigate financial, operational, and clinical incentives behind physicians' test-ordering behavior. Our model reveals that insurance coverage is a key driver behind overtesting, and the copayment and the coinsurance rate affect the equilibrium service rate in opposite ways: with a higher copayment the physician orders more tests; with a higher coinsurance rate the physician orders fewer tests. Then we show that setting a reimbursement ceiling alone cannot eliminate overtesting, and, surprisingly, overtesting can still occur even when the such a ceiling is low. Furthermore, when physicians are concerned about inaccurate diagnosis, we show that both overtesting and undertesting are possible outcomes, and the waiting time in equilibrium is shorter than the socially efficient level. We also consider two extensions. First, we consider patient heterogeneity and show that the resultant service rate becomes lower in both the market equilibrium and the social optimum. Second, we address the issue of information asymmetry about physicians' skill levels, and find that physicians' signaling efforts can lead to more salient overtesting behavior, especially when technological advancements flatten out differentiation among physicians.

In closing, we highlight a few key operational and policy implications from our work. First, overtesting is a complex phenomenon that cannot be eliminated by simple changes in the payment scheme, such as imposing a reimbursement ceiling, or eliminating insurance coverage all at once. As physicians' test-ordering behavior is closely tied to patients' strategic responses, a comprehensive understanding of physicians' and patients' financial, operational, and clinical incentives is essential before embarking on any radical changes in public policy. Second, physicians' misdiagnosis concerns lead to overtesting only when bundled together with a certain incentive environment. To address the issue of overtesting, therefore, requires not only physicians' expanding implementation of evidence-based guidelines (Walshe and Rundall 2001), but also an integrated design of health insurance structure and tort system. Third, when patients have different insurance types, patients with good insurance plans enjoy a higher coverage compared to the population average. Reducing heterogeneity in insurance of patients would help mitigate overtesting. Fourth, there is a lack of publicly accessible knowledge of physicians' skill levels, which is important in driving physicians' test-ordering behavior. Making professional evaluation for physicians more transparent to the public, through credible and accessible channels, helps reduce overtesting associated with costly signaling efforts.

Chapter 2

Contracting for On-Time Delivery in the U.S. Influenza Vaccine Supply Chain

> If you want a [flu] shot, you're gonna have to dance for it.
>
> — "Dr. Leo Spaceman", *30 Rock*, Season 3, Episode 8

2.1 Background

Behind many conventional products are myriad unconventional business challenges. When reflecting on the vaccine industry, James Matthews of Sanofi Pasteur observes, "Even though the seasonal influenza vaccine is considered a conventional vaccine by the industry, new challenges with respect to timing and availability of strains and the composition of the influenza vaccine are the rule"(Matthews 2006). A special feature of the influenza vaccine industry is that a manufacturer does not decide the design of its own product (i.e., the composition of influenza vaccine). In the United States, for example, the Vaccine and Related Biologic Products Advisory Committee (hereafter, Committee), which is independent of manufacturers, makes recommendation to the Food and Drug Administration (FDA) about the annual vaccine composition in February or March of each year for the upcoming flu season that begins the following October. The timing of this decision creates remarkable challenges: On one hand, the production process is complex and highly uncertain; on the other hand, there is a tight time window left between the announcement of the composition and the start of the flu season. These challenges make it extremely difficult to match supply with demand; in particular, supply shortage can occur even when the total supply is abundant. As an illustrative example, influenza coverage recorded a decline to 41% in the 2000–2001 influenza season, compared to 57% in the previous season; in the meanwhile, 7.5 million vaccine doses, or 10.6% of the total supply, remained unused by the end of the season (Nowalk et al. 2005; O'Mara et al. 2003). Fukuda et al. (2002) explain this seemingly paradoxical situation as follows: "The availability of influenza vaccine [in 2000 and 2001] was significantly lower during [October and November] than in previous years, which left many clinicians and patients unable to find vaccine and led to the cancellation of many vaccination campaigns. Ironically, in both years, increasing supplies of vaccine became available in December, but the waning levels of demand resulted in substantial surpluses of unused vaccine."

The common practice that vaccine manufacturers have adopted to improve their delivery performance is to start producing vaccines prior to the Committee's announcement of the vaccine composition (Committee 2007, pp. 102–103). This option, however, involves the risk that a manufacturer's projected composition may differ from the Committee's decision—in this case, the whole batch of vaccine strains in production will have to be discarded. While the manufacturer bears the entire risk associated with this option, its benefit mostly accrues to a retailer (i.e., a health care provider) because the retailer can sell more vaccines delivered on-time by the manufacturer. Thus, a well-designed supply contract needs to provide proper incentives for the manufacturer to improve its delivery performance.

Currently, most vaccine manufacturers distribute their products through two channels, each representing roughly 50% of vaccine sales/distribution (Health Industry Distributors Association 2011). In the first channel, the manufacturers sell vaccines directly to retailers (pharmacies, hospitals, public agencies, etc.); while in the second channel, they distribute vaccines through distributors who in turn deliver vaccines to their customers (primarily small physician offices). The focus of this paper is on the first channel, for which supply

Table 2.1: Sample Contract Terms

M	Season	Contract terms
A	2010–2011	A proportion of unused doses can be returned for full credit: • doses shipped before October 15: up to 25% of the doses; • doses shipped after October 15: up to 50% of the doses.
	2009–2010	The same as in the 2010–2011 season.
	2008–2009	The same as in the 2010–2011 season except that the cut-off date is November 15.
B	2010–2011	No returns are allowed; no rebate for late-delivered items.
	2009–2010	A 10% rebate is provided for orders shipped after September 30.
	2008–2009	No returns are allowed; no rebate for late-delivered items.

contracts between manufacturers and retailers are typically signed in January for vaccines to be delivered for the next flu season starting in October. Table 2.1 provides a representative sample of the contract terms used by two major influenza vaccine manufacturers (referred to as A and B, respectively) in the U.S. market during the three consecutive influenza seasons since 2008. For instance, we may refer to the contract used by Manufacturer A as the "Delivery-Time-Dependent Quantity Flexibility (D-QF) contract," while the contract used by Manufacturer B during the 2009-2010 season is the so-called "Late-Rebate (LR) contract." It is interesting to observe that the D-QF, although new to the literature, resembles the Quantity Flexibility (QF) contract (cf. Cachon 2003) but differs in that the maximum returning quantity depends on the timing of delivery.[1]

There are a couple of observations from Table 2.1. First, the two manufacturers have used different contract terms; second, even for the same manufacturer, the contract terms may vary across different years. It seems that the industry has been experimenting different contract terms in order to improve supply chain efficiency. This indicates that there is a need to deepen our understanding about various contracts. In light of these observations, one may raise several natural questions about how to manage the influenza vaccine supply chain: Why do firms (e.g., Manufacturer A) use a complex contract (e.g., a D-QF contract) instead of its simpler counterpart (e.g., a QF contract)? How do the traditional

[1] This information is based on the actual contracts which are not publicly available. Fictitious company names are used to maintain anonymity. In addition, we have found that two other manufacturers (referred to as C and D, respectively) used the same types of contracts during the 2009-2010 season: specifically, Manufacturer C used the same contract term as that of Manufacturer A except that it allowed 100% returns only for late-delivered items, and Manufacturer D used the exact same contract term as Manufacturer B.

contracts such as the wholesale price contract and the revenue sharing contract perform? What contracting options should be recommended to improve the supply chain efficiency?

To answer these questions, we develop an analytical model that captures the following key sources of uncertainties in this supply chain:

- The product *design* is exogenous to a manufacturer because the Committee determines the composition of influenza vaccine (see above). Thus, if the manufacturer begins its production prior to design freeze (i.e., the Committee's composition decision), then it faces the risk associated with product design.
- The *delivery* lead time required for manufacturing and distributing vaccines is long (usually 6–8 months) and uncertain. Due to the long and complex processes of production, testing, releasing and distribution, a manufacturer has to make its production decision way in advance of the demand season, but its delivery of vaccine can still be delayed, especially when it begins its production after the design uncertainty is resolved.
- The *demand* is time-sensitive and uncertain. CDC (2011a) notes the time-sensitivity of the demand as follows: "manufacturers with vaccine coming off the production line in middle or late November or later may not be able to sell it all and providers receiving vaccine in this same time frame may not be able to convince patients to receive it, even though late season vaccination is encouraged and in most years will be beneficial." Moreover, according to Julie Gerberding, CDC director, "We're discovering now that the demand [for flu vaccine] is also very unpredictable" (Williams 2005). The primary reason for such uncertainty is the difficulty of predicting flu activities. During 2009-10 season, 43% of children (6 month–17 years) were vaccinated with seasonal flu vaccines, but 55% of them were vaccinated with seasonal and/or H1N1 supplement vaccine after the outbreak of H1N1 flu (Singleton 2011). Because a retailer signs a contract usually in January, its forecast for the demand for the next flu season that starts the following October is fundamentally uncertain.

By capturing the three sources of uncertainties, our model enables us to evaluate the performance of various supply contracts (see Table 2.2 for their abbreviations) and to address our research questions. We show that various well-studied contracts, such as the revenue-sharing contract, are not effective in inducing satisfactory delivery performance

Table 2.2: Abbreviation of Supply Contracts

Abbreviation	Contract
BLR	Buyback-and-Late-Rebate contract
D-QF	Delivery-time-dependent Quantity Flexibility contract
QF	Quantity Flexibility contract
LR	Late-Rebate contract

due to a *negative feedback loop* in the firms' incentives: On one hand, since the benefit of on-time delivery mostly accrues to the retailer, the manufacturer lacks the motivation to improve the on-time delivery performance, which leads to potential loss in demand. On the other hand, the concern of lost demand incentivizes the retailer to order a low quantity, which further reduces the manufacturer's incentive to improve the delivery performance of the supply chain. Hence the misaligned incentives lock the supply chain into a vicious cycle of eroding delivery performance. Specifically, we first prove that the revenue sharing contract often leads to zero delivery-improving effort from the manufacturer, and can perform even worse than the wholesale price contract. Second, we introduce two complex contracts, namely the D-QF and BLR contracts, which break this vicious cycle and coordinate the supply chain. Both of these contracts are complex contracts that have not been reported in the literature. Third, in the U.S. influenza vaccine market, the D-QF contract—instead of its simpler counterpart, the QF contract—has been adopted by some vaccine manufacturers (e.g., Manufacturers A and C), while the BLR contract has not appeared in our contract samples; however, we do observe the (irregular) use of the LR contract by Manufacturers B and D. Given the resemblance between the QF and D-QF contracts, as well as that between the LR and BLR contracts, we evaluate the performance of the QF and LR contracts against their complex counterparts. We show that these simple contracts can achieve near-optimal and robust performance under different conditions: the QF contract works well when the product margin is either very high or very low; the LR contract works well only when the retailer has a dominant bargaining power. The analysis of the above complex and simple contracts subsequently allows us to evaluate the contracts used currently in the industry, and to provide practical recommendations.

Although our study is motivated by the influenza vaccine industry, our model and analysis may apply to other industries of similar characteristics. For example, according to Jackson (2009), in the apparel and footwear industry "the trends shift quickly and unpredictably, the deadlines are short and there are harsh consequences for delays," and on-time delivery is a crucial success factor in this industry because "apparel and footwear buying cycles are transitioning into shorter seasons. The right product must be on the rack to be sold in the right season." The apparel and footwear manufacturers, often located overseas, may start producing certain products early to ensure on-time delivery, but the design or style may not be on trend for the selling season; or the manufacturers may postpone their production to observe the trend, at the risk of causing delayed deliveries. These trade-offs resemble those in the vaccine industry, and can be addressed by the same analytical framework.

2.2 Literature Review

This work draws on and contributes to the following two streams of literature. First, we contribute to the rich literature of supply contracts by evaluating various (well-known and new) types of contracts in the new environment where a supply chain faces uncertainties in design, delivery, and demand. Second, this is the first paper that studies a contracting problem between an influenza vaccine manufacturer and a health care provider based on the real contracts used in practice.

Supply contracts have been studied extensively. Below we review only the papers that are most related to this paper, while referring readers to Cachon (2003) for a comprehensive review of early work. Our paper is related to the papers that study QF contract, buyback/returns, and rebates, which are intended to mitigate *demand risk*. Durango-Cohen and Yano (2006) provide a thorough review of the QF contract and its variants. In particular, Tsay (1999) is the first to model and analyze the QF contract. He characterizes the impact of QF contract parameters on the production and forecasting decisions of a supplier and a buyer in a supply chain with uncertain demand. The buyback, revenue sharing, and return contracts have been widely studied in the literature, including Pasternack (1985), Padmanabhan and Png (1997), Lariviere (1998), Arya and Mittendorf (2004), Cachon and Lariviere (2005), and Ha and Tong (2008). Several papers in the literature have considered complex contracts that combine different aspects of previously known contracts: for example, price protection, midlife returns, and end-of-life returns (Taylor 2001), rebate and returns (Taylor 2002), price and quantity commitment (Taylor and Plambeck 2007). While these papers focus on demand uncertainty, our paper also addresses uncertainties in delivery timing and product design.

The issue of *on-time delivery* has been studied from various angles. Grout and Christy (1993) study purchasing contracts in a just-in-time setting where the delivery performance is controlled by the supplier. Under delivery uncertainty only, they show that a bonus scheme improves the on-time delivery performance. Cachon and Zhang (2006) consider the sourcing problem of a buyer whose operating costs are affected by both the procurement price and delivery lead time. They characterize the optimal procurement mechanisms and identify two simpler but effective strategies. Hwang et al. (2012) consider the per-unit penalty contract adopted by retailers (e.g., Walmart) that resembles the late-rebate (LR) contract studied in our paper, but in their model the supplier has a single production mode. Our paper differs significantly from and thus enriches the on-time delivery literature in that, in the context of the vaccine industry, there is a delivery-design tradeoff.

To mitigate delivery risk, a manufacturer in our model operates under a dual-production mode, which resembles the setting of Donohue (2000) that models a fast fashion supply chain with two production modes: one is cheap but has a long lead time, and another is expensive but more responsive to market demand. Three key differences separate our paper from Donohue's. First, in our paper, early production helps improve delivery performance, whereas in Donohue's case it reduces the production cost. Second, we consider uncertainties in design, delivery and demand, while Donohue considers only demand uncertainty. Third, our paper analyzes various supply contracts observed in the vaccine industry, while Donohue focuses on the wholesale price contract.

The second related literature studies various operational issues in the influenza vaccine supply chain. Chick et al. (2008) propose the first influenza vaccine supply chain model, and show that if a central government can select a fraction of a population to vaccinate, then the government can use a variation of the cost-sharing contract to induce a manufacturer to produce socially-optimal quantity. Deo and Corbett (2009) analyze the effect of yield uncertainty on competition among vaccine manufacturers. Cho (2010) studies the Committee's problem of choosing an optimal vaccine composition with dynamic information updating, taking into account its impact on subsequent production decisions. Arifoğlu et al. (2012) study the impact of yield uncertainty and self-interested consumers on the inefficiency in this supply chain, and analyze the effectiveness of government interventions through partial centralization. Mamani et al. (2012) and Adida et al. (2011) study how the government can induce a socially optimal vaccine coverage through subsidies to a manufacturer and consumers. Recently, Chick et al. (2012) extend Chick et al. (2008) by considering a setting where the manufacturer has to satisfy the exact demand determined by the government: if a low production yield leads to a shortfall, then the manufacturer is required to make up the difference at a higher production cost, and the government incurs an extra administrative expense. In this setting, they show that the supply chain is coordinated when the additional expense of the government is transferred to the manufacturer.[2]

Our paper makes the following contributions to this literature. First, our paper is the first to consider a healthcare provider in the U.S. influenza vaccine supply chain who places an order to a manufacturer and then distributes vaccines to consumers. This research perspective is shared by a recent case study of Deo et al. (2012). Based on the *real* contracts used in practice, we study various sophisticated contracts between a healthcare provider and a manufacturer. Second, we employ several new modeling elements that reflect the industry characteristics such as uncertain delivery timing, early production mode associated with design risks, and time-sensitive uncertain demand. Finally, while the previous literature studies the effectiveness of *potential* government interventions

[2] Chick et al. (2012) also consider the case where the manufacturer knows the production yield but the government does not (with some probability for high or low yield), and show that a menu of contracts can coordinate the supply chain. Different from Chick et al. (2012), we study the contract between a manufacturer and a healthcare provider that is signed usually in January. At that time, a production yield is highly uncertain and the manufacturer does not have much informational advantage because: (i) the growth characteristic of virus strains, which has an impact on yield, is public information (see, e.g., Committee 2007), and (ii) the manufacturer learns the yield over the course of *later* production, and is required to report any yield problem to the FDA. Similarly, a retailer does not have informational advantage about its demand because: (i) the contract is signed much in advance of the upcoming flu season, and (ii) flu activities that have an impact on the demand are publicly available on the CDC website.

through partial centralization or subsidies, we shed light on improvement opportunities through coordinating contracts between firms in this supply chain, which do not require government interventions.

2.3 Modeling Framework

We consider a supply chain consisting of two risk-neutral firms, a manufacturer and a retailer. As commonly assumed in the literature (e.g., Chick et al. 2008, Arifoğlu et al. 2012), the retailer sells a product at a fixed price p. The associated demand, denoted by ξ, follows a distribution $F(\cdot)$ with density $f(\cdot)$.[3] To model the time-sensitivity of the demand, we consider two selling periods: an *ideal* period, and a *late* period. Demand arrives during the ideal period. If, by the end of the ideal period, there is unmet demand due to inadequate supply, then a proportion $\gamma \in (0,1)$ of the unserved customers will not return for vaccination, and the rest will return during the late period. Thus, the parameter γ captures the time-sensitivity of demand.[4] Based on these demand characteristics, the retailer determines the order quantity Q, which incurs an administrative cost of $c_o(\geq 0)$ per unit that captures the burden of placing, tracking, and receiving the order.

The manufacturer operates in dual production modes: "regular" and "early", with respective subscripts r and e. Under the *regular production mode*, the manufacturer has an uncertain delivery lead time and cannot always deliver the product in a timely fashion. With probability $\alpha \in (0,1)$, the delivery is on time (i.e., satisfying the demand during the ideal period); with probability $(1-\alpha)$, the delivery is late (i.e., satisfying the demand during the late period). The manufacturer also has the *early production mode*, in which the manufac-

[3] We abstract away from modeling the detailed epidemiology and consumers' vaccination decisions in the demand forming process. In general, it is very difficult to estimate the next flu-season demand starting October based on epidemic data till January, especially because the estimation of parameters in epidemic models that capture seasonal trends over summer is extremely difficult (Earn et al. 2002, Lofgren et al. 2007). Since our model requires the local demand of a retailer, the use of a general demand distribution is sufficiently general for our analysis. In practice, the retailer we interviewed simply places an order using her forecast based on the previous year demand.

[4] Our qualitative results will not change if the value of γ depends on the quantity delivered on time, reflecting the observation that some consumers' vaccine-seeking behavior may depend on the vaccination and thus infection conditions of the population in the ideal period (see Arifoğlu et al. 2012). Specifically, we can prove that the optimal parameters of the BLR and LR contracts remain the same; see the remarks in the proofs of Propositions 2.3 and 2.5. In addition, we can numerically show that our major results regarding other contracts remain valid.

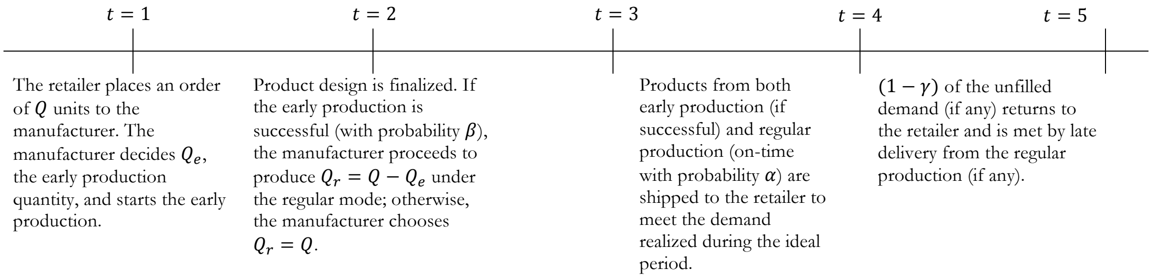
turer starts production before the product design (i.e., vaccine composition) is finalized. The early production mode guarantees on-time delivery but is vulnerable to design uncertainty: with probability $\beta \in (0,1)$, the early production uses the same composition as the finalized one, and with probability $(1-\beta)$, the early production uses a different composition, in which case the whole batch of vaccine in production has to be discarded. The respective unit production costs under the regular and early production modes are c_r and c_e. Under the dual-production-mode setting, the manufacturer first needs to decide the early production quantity, denoted by Q_e. Hence, given Q and Q_e, the regular production quantity, denoted by Q_r, can be one of following two quantities depending on the outcome of the early production mode:[5]

$$Q_r = \begin{cases} Q - Q_e & \text{with probability } \beta \\ Q & \text{with probability } (1-\beta). \end{cases} \tag{2.1}$$

Below we specify the sequence of events (see Figure 2.1):

- $t = 1$: The retailer determines an order quantity Q, and the manufacturer, after receiving the retailer order, determines an early production quantity Q_e.
- $t = 2$: Upon the release of the finalized product design, the manufacturer determines the regular production quantity Q_r according to equation (2.1).
- The ideal period (between $t = 3$ and $t = 4$): With probability β, early-production outputs are delivered to the retailer during this period. In addition, with probability α, regular-production outputs are delivered during this period.
- The late period (between $t = 4$ and $t = 5$): Among the unserved customers in the ideal period, a fraction $(1-\gamma)$ of them will return to seek vaccination. On the supply side, with probability $(1-\alpha)$, regular production outputs are delivered to satisfy such residual demand.

[5] For tractability, we adopt the *forced compliance regime* under which the manufacturer must supply Q as ordered by the retailer (cf. Cachon and Lariviere 2001), while it does not guarantee on-time delivery and hence can cause lost sales to the retailer. Chick et al. (2012) make a similar assumption, but in their model, late delivery does not cause lost sales under the deterministic demand.



The retailer places an order of Q units to the manufacturer. The manufacturer decides Q_e, the early production quantity, and starts the early production.

Product design is finalized. If the early production is successful (with probability β), the manufacturer proceeds to produce $Q_r = Q - Q_e$ under the regular mode; otherwise, the manufacturer chooses $Q_r = Q$.

Products from both early production (if successful) and regular production (on-time with probability α) are shipped to the retailer to meet the demand realized during the ideal period.

$(1 - \gamma)$ of the unfilled demand (if any) returns to the retailer and is met by late delivery from the regular production (if any).

Figure 2.1: Sequence of events in the dual-production-mode setting

For notational convenience, we define an indicator I to be

$$I = \begin{cases} 1 & \text{if products from the regular production are delivered on time,} \\ 0 & \text{otherwise.} \end{cases}$$

Similarly, we define another indicator $J \in \{1, 0\}$ to represent whether or not the early production matches the Committee's recommendation.[6] We have $\Pr(I = 1) = \alpha$, $\Pr(I = 0) = 1 - \alpha, \Pr(J = 1) = \beta$, and $\Pr(J = 0) = 1 - \beta$. Table **??** lists three cases depending on the realization of I and J.

Table 2.3: Three Cases of Delivery and Early Production Outcome

Case 1.	$I = 1$ (on time delivery)	with probability α
Case 2.	$I = 0$ (late delivery) and $J = 1$ (successful early production)	with probability $(1 - \alpha)\beta$
Case 3.	$I = 0$ (late delivery) and $J = 0$ (unsuccessful early production)	with probability $(1 - \alpha)(1 - \beta)$

[6] The uncertainty associated with the production of influenza vaccine has two dimensions: delivery timing and output quantity. As discussed earlier, the first dimension has been one of the primary causes of mismatch between demand and supply, but has never been studied in the literature. Thus, our paper focuses on the first dimension, while suppressing the second dimension to maintain tractability. If we incorporate the second dimension as the fourth source of uncertainty in our model, then we need to model a manufacturer's choice of the quantity of raw materials (i.e., chicken eggs) for both early and regular production modes (denoted by n_e and n_r, respectively). If we assume, using Proposition 1 in Chick et al. (2008), that the manufacturer sets the total quantity of raw materials as Q/K (so that $n_r = Q/K - J \cdot n_e$), where $K > 0$ is a constant determined by the distribution of the random yield, then all of our results remain valid.

The *ex post* sales quantity, denoted by Z, falls into one of the following three cases:

$$Z(Q, Q_e|\xi) = \begin{cases} \min\{Q, \xi\} & \text{(Case 1)} \\ \min\{Q_e, \xi\} + \min\{Q - Q_e, (1-\gamma)(\xi - Q_e)^+\} & \text{(Case 2)} \\ \min\{Q, (1-\gamma)\xi\} & \text{(Case 3),} \end{cases} \tag{2.2}$$

where $(\xi - Q_e)^+ = \max\{\xi - Q_e, 0\}$. In Case 1, those items produced in the regular mode are delivered on-time, and thus the whole order Q is delivered on-time, regardless of the outcome of the early production; in Case 2, the regular production is late and the early production is successful; and in Case 3, the regular production is late and the early production is not successful. The expected sales quantity is thus:

$$\begin{aligned} E[Z(Q, Q_e)] = \alpha&\left[\int_0^Q \xi dF(\xi) + \int_Q^\infty Q dF(\xi)\right] \\ &+ (1-\alpha)\beta\left\{\int_0^{Q_e} \xi dF(\xi) + \int_{Q_e}^{\frac{Q-\gamma Q_e}{1-\gamma}} [\gamma Q_e + (1-\gamma)\xi] dF(\xi) + \int_{\frac{Q-\gamma Q_e}{1-\gamma}}^\infty Q dF(\xi)\right\} \\ &+ (1-\alpha)(1-\beta)\left[\int_0^{\frac{Q}{1-\gamma}} (1-\gamma)\xi dF(\xi) + \int_{\frac{Q}{1-\gamma}}^\infty Q dF(\xi)\right]. \end{aligned} \tag{2.3}$$

Note that Q_e appears only in the second term of the right-hand side of (2.3), suggesting that the early production helps ease the supply shortage when the regular production mode yields late delivery.

For future comparison, we now analyze the first-best scenario in which a central decision maker jointly determines the order quantity Q and the early production quantity Q_e to maximize the supply chain profit. Let $\pi_S(Q, Q_e)$ denote the expected profit of the supply chain. The solution to the following problem:

$$\max_{Q>0, 0\leq Q_e \leq Q} \pi_S(Q, Q_e) = pE[Z(Q, Q_e)] - [(c_r + c_o)Q + (c_e - \beta c_r)Q_e] \tag{2.4}$$

is referred to as the first-best solution and denoted by (Q^{FB}, Q_e^{FB}). In (2.4), $(c_r + c_o)Q + (c_e - \beta c_r)Q_e$ is the total expected cost, and is derived from

$$c_eQ_e + c_rE[Q_r] + c_oQ = c_eQ_e + c_r[\underbrace{\beta(Q - Q_e)}_{J=1} + \underbrace{(1-\beta)Q}_{J=0}] + c_oQ.$$

We assume that $c_e > \beta c_r$—which includes the case that $c_e = c_r$—to focus on a realistic situation where there is a tradeoff between the delivery advantage of early production and informational advantage of regular production. This assumption guarantees the following result:

Lemma 2.1 $Q_e^{FB} < Q^{FB}$.

All proofs are presented in the appendices. In the first-best solution, it is possible that the manufacturer operates only under the regular production mode, i.e., $Q_e^{FB} = 0$. Indeed, we can show that it is worthwhile introducing the early production mode only when the delivery from the regular production mode is unreliable (i.e., α is low), the prediction for the finalized design is sufficiently accurate (i.e., β is high), and the early production mode is inexpensive (i.e., $c_e - \beta c_r$ is low).

We proceed to consider a decentralized supply chain under different supply contracts. A contract is said to *coordinate* the supply chain if it induces the first-best solution (Q^{FB}, Q_e^{FB}) from the firms comprising the supply chain. One major performance metric we use in evaluating supply contracts is the efficiency of the supply chain, which is defined as the ratio of the supply chain's expected profit under a contract to that in the first-best scenario. We focus on the interesting case where $Q_e^{FB} > 0$ because $Q_e^{FB} = 0$ means that the early production is not a viable option. We start with analyzing the revenue sharing contract in §2.4, followed by two complex contracts in §2.5, and then compare the performance with their simpler counterparts in §2.6–2.7.

2.4 Impact of Negative Incentive Feedback Loop

A number of coordinating contracts have been proposed and studied in the supply chain management literature (e.g., Cachon 2003). Thus, we start with analyzing these existing coordination contracts in our setting. In this section, we first analyze the revenue sharing (RS) contract, and then introduce the notion of the negative feedback loop formed between the two firms. Under this contract, the retailer pays the manufacturer a wholesale price w for each unit of order, plus a percentage, denoted by ψ, of the retailer's revenue. Given $\psi \in (0,1)$, we can express the manufacturer's profit and the retailer's profit, respectively, as follows:

$$\pi_M^{RS}(Q, Q_e) = \psi p E[Z(Q, Q_e)] + (w - c_r)Q - (c_e - \beta c_r)Q_e, \text{ and}$$
$$\pi_R^{RS}(Q, Q_e) = (1 - \psi) p E[Z(Q, Q_e)] - (w + c_o)Q.$$

It is well-known that revenue sharing achieves coordination for a traditional supply chain consisting of a newsvendor retailer and a supplier (see, e.g., Cachon and Lariviere 2005). Does this result continue to hold in our supply chain that involves three types of uncertainties? The next proposition provides the answer to this question.

Proposition 2.1 *If $(1 - \alpha)\beta\gamma[p - (w + c_o)] < c_e - \beta c_r$, then*

(i) Under a revenue sharing contract, the manufacturer will always choose Q_e=0.

(ii) The revenue sharing contract is strictly dominated by the wholesale price contract in terms of supply chain efficiency.

Proposition 2.1 is a surprising result. First, the revenue sharing contract fails to coordinate the supply chain because it induces $Q_e = 0$; second, the supply chain performance under the revenue sharing contract is even worse than that under the wholesale price contract because under the revenue sharing contract the retailer orders even less than under the wholesale price contract. This indicates that the existing results in the literature (e.g., revenue sharing can coordinate a supply chain) may no longer hold in our problem setting. We offer the following explanation to this counter-intuitive result.

A prominent feature of the vaccine supply chain is the presence of the delivery uncertainty. One way to mitigate this uncertainty is to use the early production mode. However, early production is risky because of the product design uncertainty, and, as a result, the manufacturer may not have the right incentive when choosing the early production quantity. Thus, compared to the traditional supply chain setting, in addition to the retailer's ordering decision, we also need to coordinate the manufacturer's early production decision. Clearly, from the result that $Q_e = 0$ in Proposition 2.1, we know that the revenue sharing scheme does not resolve the incentive problem for the manufacturer at all. This, in conjunction with the observation that the retailer orders less under the revenue sharing contract than under the wholesale price contract, implies that revenue sharing provides even less incentive to overcome double marginalization than the wholesale price contract. This is in stark contrast with the supply chain coordination literature where it is widely believed that revenue sharing can induce a higher order quantity from the retailer.

Close scrutiny of the supply chain dynamics may help us reveal the driving force behind this surprising result. To this end, we first characterize the relationship between the two critical decisions in the supply chain (which holds irrespective of a specific contract used).

Lemma 2.2 *(i)* $\partial E[Z(Q, Q_e)]/\partial Q_e \geq 0$, $\partial E[Z(Q, Q_e)]/\partial Q > 0$.

(ii) $\partial^2 E[Z(Q, Q_e)]/(\partial Q \partial Q_e) > 0$.

Lemma 2.2(i) means that the expected sales $E[Z(Q, Q_e)]$ increases both in the manufacturer's early production quantity Q_e and in the retailer's order quantity Q. Lemma 2.2(ii) suggests that the marginal benefit of a higher order quantity Q increases in Q_e, because the retailer can expect more on-time delivered products from increased Q_e and hence a lower chance of lost sales. Interestingly, Lemma 2.2(ii) also suggests that the marginal benefit of early production increases in the order quantity Q as well. In other words, the decisions Q_e and Q are complementary to each other: a higher order quantity Q will make early production more beneficial to the supply chain, and vice versa. This leads to a negative feedback loop in the firms' incentives in our supply chain: On one hand, when the manufacturer bears the risk associated with the early production mode, it lacks the motivation to improve the on-time delivery performance, which leads to potential loss in demand; on the other hand, the demand loss incentivizes the retailer to reduce its order quantity, which further discourages the manufacturer from making efforts to achieve on-time delivery.

The above vicious cycle is instrumental behind the surprising outcome presented in Proposition 2.1. Since revenue sharing offers little incentive for the manufacturer to use early production, the retailer will choose a relatively low order quantity Q; this further prevents the manufacturer from operating in the early production mode. In fact, the negative incentive is so strong that the manufacturer often chooses $Q_e = 0$ under the revenue sharing contract. In this case, since the manufacturer chooses not to operate in the early production mode, the percentage of revenue sharing (ψ) does not have any impact on the manufacturer's decision. On the retailer's side, however, a higher ψ induces the retailer to place a smaller order, hence deviating further from the first-best order quantity. Therefore, revenue sharing may induce an even lower order quantity than the wholesale price contract, which implies an inferior supply chain performance.

To understand the condition given in Proposition 2.1 (i.e., $(1-\alpha)\beta\gamma[p-(w+c_o)] < c_e - \beta c_r$), note that in its left-hand side, $\gamma[p-(w+c_o)]$ represents the retailer's maximum proportion (γ) of lost net revenue $[p-(w+c_o)]$ when late delivery occurs, and $(1-\alpha)\beta$ represents the probability for the early production mode to effectively function against late delivery (cf. Case 2 in Table **??**). Hence the left-hand side is the maximum expected benefit of the earlier production mode to the retailer. The right-hand side of the condition is the expected additional cost of operating in the early production mode. The condition suggests that, even if the retailer is willing to share all the benefits of on-time delivery with the manufacturer, the manufacturer still lacks the incentive to operate in the early production mode. When the condition is violated, we can numerically show that the manufacturer's chosen early production quantity is still substantially lower than the first-best level, and that the revenue sharing contract does not perform significantly better than the wholesale price contract. To our knowledge, the above result (i.e., the wholesale price contract can outperform the revenue sharing contract) has not been previously reported in the literature. Nevertheless, it seems to corroborate the industry observation: While we are unaware of the revenue sharing arrangement, we do observe several instances of the wholesale price contract in the influenza vaccine industry (e.g., Manufacturer B in Table 2.1).

Similarly, we have examined other coordinating contracts studied commonly in the literature (such as buyback, quantity flexibility, quantity discount, and sales rebate), and find that they cannot coordinate our supply chain, either. To save space, we present the detailed analysis of these contracts in the appendices. A common problem in these contracts is that they do not provide sufficient incentives for the manufacturer to choose the first-best early production quantity; hence, due to the negative feedback loop described above, the retailer will not use the first-best order quantity, either. In the next section we propose two contracts that can overcome this negative feedback loop and coordinate the supply chain.

2.5 Supply Contracts for Channel Coordination

So far we have shown that due to the negative feedback loop, traditional contracts cannot coordinate the vaccine supply chain. In this section, we apply two criteria when searching for coordinating contracts: First, the contract must be able to induce the desirable actions from the supply chain firms (i.e., they are indeed coordinating); second, they are connected to the real contracts observed in practice (i.e., they are practically appealing). Also, we assume that the manufacturer's early production quantity is not verifiable by the retailer and hence cannot be contracted. We propose two coordinating contracts: D-QF contract in §2.5.1, and BLR contract in §2.5.2.

2.5.1 Delivery-Time-Dependent Quantity Flexibility (D-QF) Contract

Under the traditional QF contract, the retailer is allowed to return no more than a portion of the order quantity at full price. By contrast, under a D-QF contract, the maximum returning quantity (hereafter return allowance) depends on the timing of delivery. To our best knowledge, while this contract has been adopted by some vaccine manufacturers (see Manufacturer A in Table 2.1), it has not been reported in the literature. Let Y_1 be the shipping quantity by the end of the ideal period, and Y_2 the shipping quantity after the ideal period. The return allowance is then equal to $\kappa_1 Y_1 + \kappa_2 Y_2$, where $\kappa_1, \kappa_2 \in [0, 1]$ are the returnable proportions of the delivery quantities for on-time and late deliveries, respectively.

Hence the return allowance is a random variable and can be represented as

$$\kappa_1 Y_1 + \kappa_2 Y_2 = \begin{cases} \kappa_1 Q & \text{(Case 1)} \\ \kappa_1 Q_e + \kappa_2 (Q - Q_e) & \text{(Case 2)} \\ \kappa_2 Q & \text{(Case 3).} \end{cases} \tag{2.5}$$

When $\kappa_1 = \kappa_2$, such a return policy is equivalent to the QF contract previously studied in the literature. We denote by $R_d(Q, Q_e)$ the total returning quantity at the end of a demand season under the D-QF contract: $R_d(Q, Q_e) = \min\{\kappa_1 Y_1 + \kappa_2 Y_2, Q - Z(Q, Q_e)\}$. The transfer payment from the manufacturer to the retailer is given by $T_d(Q, Q_e) = w \cdot R_d(Q, Q_e)$. Comparing the return allowance $\kappa_1 Y_1 + \kappa_2 Y_2$ with the leftover inventory $Q - Z(Q, Q_e)$, we can express the returning quantity $R_d(Q, Q_e)$ as follows.

Case 1. $I = 1$. In this case, $\kappa_1 Y_1 + \kappa_2 Y_2 = \kappa_1 Q$ from (2.5) and $Q - Z(Q, Q_e) = Q - \min\{Q, \xi\}$ from (2.2). Therefore,

$$R_d(Q, Q_e | I = 1) = \begin{cases} \kappa_1 Q & \text{if } \xi < (1 - \kappa_1) Q \\ Q - \xi & \text{if } (1 - \kappa_1) Q \le \xi < Q \\ 0 & \text{if } \xi \ge Q. \end{cases} \tag{2.6}$$

Case 2. $I = 0$ and $J = 1$. In this case, $\kappa_1 Y_1 + \kappa_2 Y_2 = \kappa_1 Q_e + \kappa_2 (Q - Q_e)$ from (2.5), and $Q - Z(Q, Q_e) = Q - \min\{Q_e, \xi\} - \min\{Q - Q_e, (1 - \gamma)(\xi - Q_e)^+\}$ from (2.2). For ease of exposition, let us define the following three numbers:

$\xi^{(1)} \equiv (1 - \kappa_2) Q + (\kappa_2 - \kappa_1) Q_e$ so that $\xi \ge \xi^{(1)} \Leftrightarrow Q - \xi \le \kappa_1 Q_e + \kappa_2 (Q - Q_e)$;

$\xi^{(2)} \equiv [(1 - \kappa_2) Q + (\kappa_2 - \kappa_1 - \gamma) Q_e]/(1 - \gamma)$ so that $\xi \ge \xi^{(2)} \Leftrightarrow Q - \{Q_e + (1 - \gamma)(\xi - Q_e)\} = Q - \gamma Q_e - (1 - \gamma)\xi \le \kappa_1 Q_e + \kappa_2 (Q - Q_e)$;

$\xi^{(3)} \equiv (Q - \gamma Q_e)/(1 - \gamma)$ so that $\xi \ge \xi^{(3)} \Leftrightarrow Q - Q_e \le (1 - \gamma)(\xi - Q_e)$;

where $\xi^{(1)}$ (resp., $\xi^{(2)}$) is the demand level at which the leftover inventory is equal to the return allowance when $\xi < Q_e$ (resp, $Q_e \le \xi < \xi^{(3)}$), and $\xi^{(3)}$ is the demand level above which there will be no leftover inventory (implying that $\xi^{(2)} < \xi^{(3)}$). By comparing the leftover inventory with the return allowance, we can derive the return quantity in each of the following two subcases:

(a) If $\kappa_2 - \kappa_1 > 1 - (1-\kappa_2)Q/Q_e$ ($\Leftrightarrow \xi^{(1)} > Q_e$), then we can show $\xi^{(2)} > Q_e$, so that $R_d(Q, Q_e|I = 0, J = 1)$ becomes:

$$\begin{cases} \min\{Q - \xi, \kappa_1 Q_e + \kappa_2(Q - Q_e)\} = \kappa_1 Q_e + \kappa_2(Q - Q_e) & \text{if } 0 \le \xi < Q_e \\ \min\{Q - \gamma Q_e - (1-\gamma)\xi, \kappa_1 Q_e + \kappa_2(Q - Q_e)\} = \kappa_1 Q_e + \kappa_2(Q - Q_e) & \text{if } Q_e \le \xi < \xi^{(2)} \\ \min\{Q - \gamma Q_e - (1-\gamma)\xi, \kappa_1 Q_e + \kappa_2(Q - Q_e)\} = Q - \gamma Q_e - (1-\gamma)\xi & \text{if } \xi^{(2)} \le \xi < \xi^{(3)} \\ 0 & \text{if } \xi \ge \xi^{(3)}. \end{cases} \tag{2.7}$$

(b) If $\kappa_2 - \kappa_1 \le 1 - (1-\kappa_2)Q/Q_e \Leftrightarrow (\xi^{(1)} \le Q_e)$, then we can show $\xi^{(2)} \le Q_e$, so that $R_d(Q, Q_e|I = 0, J = 1)$ becomes:

$$\begin{cases} \min\{Q - \xi, \kappa_1 Q_e + \kappa_2(Q - Q_e)\} = \kappa_1 Q_e + \kappa_2(Q - Q_e) & \text{if } 0 \le \xi < \xi^{(1)} \\ \min\{Q - \xi, \kappa_1 Q_e + \kappa_2(Q - Q_e)\} = Q - \xi & \text{if } \xi^{(1)} \le \xi < Q_e \\ \min\{Q - \gamma Q_e - (1-\gamma)\xi, \kappa_1 Q_e + \kappa_2(Q - Q_e)\} = Q - \gamma Q_e - (1-\gamma)\xi & \text{if } Q_e \le \xi < \xi^{(3)} \\ 0 & \text{if } \xi \ge \xi^{(3)}. \end{cases} \tag{2.8}$$

Case 3. $I = 0$ and $J = 0$. In this case, $\kappa_1 Y_1 + \kappa_2 Y_2 = \kappa_2 Q$ from (2.5), and $Q - Z(Q, Q_e) = Q - \min\{Q, (1-\gamma)\xi\}$ from (2.2). Therefore,

$$R_d(Q, Q_e|I = 0, J = 0) = \begin{cases} \kappa_2 Q & \text{if } 0 \le \xi < \frac{(1-\kappa_2)Q}{1-\gamma} \\ Q - (1-\gamma)\xi & \text{if } \frac{(1-\kappa_2)Q}{1-\gamma} \le \xi < \frac{Q}{1-\gamma} \\ 0 & \text{if } \xi \ge \frac{Q}{1-\gamma}. \end{cases} \tag{2.9}$$

The following proposition provides the necessary condition for the D-QF contract to coordinate the supply chain when $Q_e^{FB} > 0$. Later in section 2.6.1, we prove that the QF contract never coordinates the supply chain unless early production is not a viable option (i.e., $Q_e^{FB} = 0$).

Proposition 2.2 *When $Q_e^{FB} > 0$, to coordinate the supply chain, the D-QF contract must have parameters (κ_1, κ_2) that satisfy $\kappa_2 > \kappa_1$, and one of the following two conditions:*

(i) $\kappa_1 Q_e^{FB} + \kappa_2(Q^{FB} - Q_e^{FB}) < Q^{FB} - Q_e^{FB}$ *and*

$$(\kappa_2 - \kappa_1 - \gamma) \cdot F\left(\frac{(1-\kappa_2)Q^{FB} + (\kappa_2 - \kappa_1 - \gamma)Q_e^{FB}}{1-\gamma}\right) + \gamma F(Q_e^{FB}) = \frac{(c_e - \beta c_r)(p-w)}{wp(1-\alpha)\beta}; \quad (2.10)$$

(ii) $\kappa_1 Q_e^{FB} + \kappa_2(Q^{FB} - Q_e^{FB}) \geq Q^{FB} - Q_e^{FB}$ *and*

$$(\kappa_2 - \kappa_1) \cdot F\left((1-\kappa_2)Q^{FB} + (\kappa_2 - \kappa_1)Q_e^{FB}\right) = \frac{(c_e - \beta c_r)(p-w)}{wp(1-\alpha)\beta}. \quad (2.11)$$

The proof of the proposition is technical but its intuition is as follows. There are two contract parameters, κ_1 and κ_2, to specify in the D-QF contract. To coordinate the supply chain, a contract must not only induce the retailer to choose an order quantity of Q^{FB} units, but also induce the manufacturer to choose an early production quantity of Q_e^{FB}. The first condition in Proposition 2.2 that $\kappa_2 > \kappa_1$ is expected because otherwise the manufacturer will not bear the delivery risk associated with regular production. The condition $\kappa_2 > \kappa_1$ has also been observed in practice: in the example shown in Table 2.1, Manufacturer A uses $\kappa_2 = 50\% > \kappa_1 = 25\%$. In addition to $\kappa_2 > \kappa_1$, one of the two conditions (i) and (ii) in Proposition 2.2 must be satisfied to guarantee that the manufacturer choose Q_e^{FB} when the retailer orders Q^{FB}. The conditions (i) and (ii) correspond to the scenario that the total return allowance is low and high, respectively. This can be seen by noting that $\kappa_1 Q_e^{FB} + \kappa_2(Q^{FB} - Q_e^{FB})$ and $Q^{FB} - Q_e^{FB}$ are the return allowance and the late-delivered quantity, respectively, for Case 2 when Q^{FB} and Q_e^{FB} are chosen. In both (2.10) and (2.11), the left-hand side increases in $\kappa_2 - \kappa_1$ and measures the incentive of the D-QF contract provided to the manufacturer, while the right-hand side captures the cost of operating in early production, which increases with $c_e - \beta c_r$.

Proposition 2.2 provides the necessary conditions for the D-QF contract to coordinate the supply chain; it does not guarantee the existence of a coordinating D-QF contract. To coordinate the supply chain, the contract also needs to motivate the retailer to choose an ordering quantity that matches its first-best level. In our numerical experiments presented in §2.7, the D-QF contract coordinates the supply chain in 48.1% of the instances; in the case when the D-QF contract does not coordinate the supply chain, it still induces a fairly high supply chain efficiency with the average of 98.8%.

2.5.2 Buyback-and-Late-Rebate (BLR) Contract

The second coordinating contract, namely BLR contract, is a variant of the well-known buyback contract. Specifically, we incorporate a late rebate term into the traditional buyback contract. Note that the LR contract has been observed in the vaccine industry (see Table 1). Under the BLR contract, the manufacturer provides the retailer with a rebate for late-delivered products in addition to providing the retailer with a buyback credit for each unsold unit. This contract is based on two quantities that are observable to both the manufacturer and the retailer: the leftover inventory and the late-delivered quantity. We use $\rho \in (0,1)$ such that $\rho \cdot w$ is the rebate from the wholesale price w for late-delivered unit. Thus, the expected transfer payment from the manufacturer to the retailer is represented as

$$b \cdot \{Q - E[Z(Q, Q_e)]\} + \rho w \left[(1-\alpha)\beta(Q - Q_e) + (1-\alpha)(1-\beta)Q\right]. \tag{2.12}$$

In (2.12), the first term is the manufacture's expected buyback credit to the retailer, and the second term is the manufacturer's expected rebate, in which late-delivered quantity is 0 with probability α (in Case 1), $Q - Q_e$ with probability $(1-\alpha)\beta$ (in Case 2), and Q with probability $(1-\alpha)(1-\beta)$ (in Case 3). The proposition below details the BLR contract that coordinates the supply chain.

Proposition 2.3 *The BLR contract coordinates the supply chain if and only if*

$$b^*_{BLR} = \frac{\beta w - c_e}{\beta(p - c_o) - c_e} \cdot p, \text{ and } \rho^*_{BLR} = \frac{(p - w - c_o)(c_e - \beta c_r)}{w(1-\alpha)[\beta(p - c_o) - c_e]}.$$

The following corollary provides comparative statics to show the impact of α, β, and γ on the optimal BLR contract parameters b^*_{BLR} and ρ^*_{BLR}.

Corollary 2.1 *Under the BLR contract, the following results hold:*

	b^*_{BLR}	ρ^*_{BLR}	*R's* *profit share*	*M's* *profit share*
α	-	↑	-	-
β	↑	↓	↑	↓
γ	-	-	-	-

In the BLR contract, while the rebate component motivates the manufacturer to operate in the early production mode, the buyback component incentivizes the retailer to increase the order quantity. As α increases, the regular production mode becomes more reliable. Thus a higher rebate rate (ρ^*_{BLR}) is needed to motivate the manufacturer to operate in the early production mode, but the buyback price (b^*_{BLR}) remains unchanged because α does not affect the expected additional cost of early production ($c_e - \beta c_r$). The profit division of the supply chain between the two firms is directly tied to the buyback price b^*_{BLR} and thus remains unchanged. As β increases, early production is less costly. Even a lower rebate rate (ρ^*_{BLR}) can now provide the manufacturer with adequate incentive. To encourage the retailer to place a large order, the manufacturer offers a higher buyback price (b^*_{BLR}), which gives the retailer a higher profit share. Interestingly, the optimal contract parameters are independent of the time-sensitivity parameter γ, which does not affect the cost structure or the delivery performance of the supply chain. Thus, the optimal BLR contract coordinates the supply chain even when γ is a function of the on-time-delivered quantity (see the proof of Proposition 2.3).

We close this section by briefly discussing the flexibility of profit division between the manufacturer and the retailer. Under the optimal BLR contract, the retailer's profit share is $b^*_{BLR}/p = (\beta w - c_e)/[\beta(p - c_o) - c_e]$. Therefore, the profit of the supply chain can be arbitrarily divided between the firms by adjusting wholesale price w. Likewise, we can numerically show that under the D-QF contract, any profit division between the firms can be achieved by adjusting w.

2.6 Analysis of Simple Contracts

This section evaluates the performance of the two simpler contracts, namely, QF and LR contracts. Our analysis in this section focuses on their analytical properties, which will be complemented by numerical experiments in the next section.

2.6.1 Quantity Flexibility (QF) Contract

We have shown that the D-QF contract can conditionally coordinate the supply chain. The D-QF contract needs to specify two parameters (κ_1 and κ_2), and requires tracking the quantity of on-time- and late-delivered items. This makes the D-QF contract more cumbersome to implement in practice than a simple QF contract. Thus, we analyze the QF contract, a well-studied contract under which the manufacturer provides the retailer with full credit for the leftover inventory up to a pre-determined threshold. It is much simpler than the D-QF contract because it involves only a single parameter and does not depend on the timing of delivery.

While there exist different specifications of the threshold, we focus on the QF contract where the threshold is a proportion of the order quantity, as is commonly assumed in the QF literature (e.g., Tsay 1999, Cachon and Lariviere 2001, and Plambeck and Taylor 2005). We denote by $\kappa \in (0,1)$ the proportion of the retailer's ordering quantity Q that is allowed to return after the sales season. This contract can be viewed as a special case of the D-QF contract where $\kappa_1 = \kappa_2$. The returning quantity, denoted by $R(Q,Q_e)$, is then $R(Q,Q_e) = \min\{\kappa Q, Q - Z(Q,Q_e)\}$, where $Q - Z(Q,Q_e)$ is the leftover inventory. Since the manufacturer provides full credit for returns, its total transfer payment to the retailer, denoted by $T_c(Q,Q_e)$, is equal to $w \cdot R(Q,Q_e)$. Similar to the D-QF contract in §2.5.1, the returning quantity $R(Q,Q_e)$ takes different values, as shown in Table 2.4.

While the literature establishes that the QF contract can coordinate a supply chain when a set of conditions are met (cf. Cachon 2003, §6.2.5), the following proposition states otherwise.

Proposition 2.4 *Unless the early production mode is not a viable option (i.e., $Q_e^{FB} = 0$), the QF contract does not coordinate the supply chain.*

Table 2.4: Returning Quantity $R(Q, Q_e)$ Under Different Cases

Case 1.	$R(Q,Q_e) = \begin{cases} \kappa Q & \text{if } 0 \leq \xi < (1-\kappa)Q \\ Q - \xi & \text{if } (1-\kappa)Q \leq \xi < Q \\ 0 & \text{if } \xi \geq Q. \end{cases}$	
Case 2.	When $0 \leq \kappa < 1 - Q_e/Q$: $R(Q,Q_e) = \begin{cases} \kappa Q & \text{if } 0 \leq \xi < \frac{(1-\kappa)Q - \gamma Q_e}{1-\gamma} \\ Q - \gamma Q_e - (1-\gamma)\xi & \text{if } \frac{(1-\kappa)Q-\gamma Q_e}{1-\gamma} \leq \xi < \frac{Q-\gamma Q_e}{1-\gamma} \\ 0 & \text{if } \xi \geq \frac{Q-\gamma Q_e}{1-\gamma}. \end{cases}$	When $1 - Q_e/Q \leq \kappa \leq 1$: $R(Q,Q_e) = \begin{cases} \kappa Q & \text{if } 0 \leq \xi < (1-\kappa)Q \\ Q - \xi & \text{if } (1-\kappa)Q \leq \xi < Q_e \\ Q - \gamma Q_e - (1-\gamma)\xi & \text{if } Q_e \leq \xi < \frac{Q-\gamma Q_e}{1-\gamma} \\ 0 & \text{if } \xi \geq \frac{Q-\gamma Q_e}{1-\gamma}. \end{cases}$
Case 3.	$R(Q,Q_e) = \begin{cases} \kappa Q & \text{if } 0 \leq \xi < \frac{(1-\kappa)Q}{1-\gamma} \\ Q - (1-\gamma)\xi & \text{if } \frac{(1-\kappa)Q}{1-\gamma} \leq \xi < \frac{Q}{1-\gamma} \\ 0 \text{ if } \xi \geq \frac{Q}{1-\gamma}. \end{cases}$	

Remark 2.1 *QF contract is a special case of the D-QF contract with $\kappa_1 = \kappa_2$. If we relate this to the necessary conditions for supply chain coordination in Proposition 2.2, it is apparent that the QF contract satisfies the first part of the condition (ii), but does not satisfy the second part.*

Remark 2.2 *Like the revenue sharing contract (as discussed in §2.4), the QF contract does not coordinate the supply chain because the manufacturer is not provided with adequate incentive to operate in the early production mode. However, while revenue sharing leads to a vicious cycle of eroding delivery performance and can perform even worse than the wholesale price contract, we later demonstrate in our numerical experiments that the QF contract motivates the retailer to choose a larger order quantity, and hence does not lead to the vicious cycle.*

2.6.2 Late Rebate (LR) Contract

Our analysis in §2.5.2 suggests that the BLR contract can coordinate the supply chain. In the vaccine industry, however, we observe that manufacturers adopt a simple LR contract under which they offer a rebate for orders shipped after the ideal vaccination period. For example, Manufacturer B provided a 10% rebate for late-delivered doses during the 2009-2010 season (see Table 2.1). Compared to the BLR contract, the LR contract is easier to implement because it does not require tracking left-over inventory.

We denote by ρ_{LR} a proportion of the wholesale price w that the manufacturer rebates to the retailer for late-delivered items. Under this contract, the manufacturer's expected profit and the retailer's expected profit can be expressed respectively as follows:

$$\pi_M^{LR} = (w - c_r)Q - (c_e - \beta c_r)Q_e - \rho_{LR} w(1-\alpha)(Q - \beta Q_e) \text{ and}$$
$$\pi_R^{LR} = p \cdot E[Z(Q, Q_e)] - [w + c_o - \rho_{LR} w(1-\alpha)]Q - \rho_{LR} w\beta(1-\alpha)Q_e.$$

Next, we provide the condition for the LR contract to coordinate the supply chain.

Proposition 2.5 *The LR contract coordinates the supply chain if and only if the wholesale price $w = c_e/\beta$, and the rebate level $\rho_{LR}^* = (c_e - \beta c_r)/[(1-\alpha)c_e]$. Under this contract, the retailer fully bears the risk associated with design uncertainty, and gains all the supply chain profit.*

Proposition 2.5 shows that the LR contract coordinates the supply chain only when the retailer has a dominating bargaining power and earns all supply chain profit. This contract requires a wholesale price above the unit production cost (since $w = c_e/\beta$ is greater than both c_e and c_r), and a rebate that depends on the quantity of late-delivered items. This simple contract effectively transfers all the risk due to early production from the manufacturer to the retailer without necessitating the retailer's engagement in monitoring the manufacturer's production activities.

Since the retailer is unlikely to have a dominating bargaining power in practice, it is usually the case that $w > c_e/\beta$, and the retailer has to share its profit with the manufacturer. Thus, we evaluate the LR contract for the case when $w > c_e/\beta$ in our numerical study.

2.7 Numerical Study

In §2.6, we have analyzed the QF and LR contracts, and showed that the QF contract cannot coordinate the supply chain, while the LR contract can do so only under a restrictive condition. In §2.7.1 and §2.7.2, we numerically evaluate their performance against their complex counterparts, D-QF and BLR contracts studied in §2.5.1 and §2.5.2, respectively. This way we can shed some light on the trade-off between contract simplicity and performance. Then in §2.7.3, we evaluate the contracts listed in Table 2.1 and provide recommendations for improving the supply chain efficiency.

As is common in the supply contract literature, the demand is assumed to follow a Gamma distribution, with a density of $f(\xi) = b^{-a}e^{-\xi/b}\xi^{a-1}/\Gamma(a)$, and a cumulative density of $F(\xi) = \gamma\left(a, \xi/b\right)/\Gamma(a), \xi > 0$. In practice, we have observed manufacturers use the same contracts for different retailers. Because we consider a retailer of any size, we normalize the mean to 1 and use a coefficient of variation of 0.5; this leads to the shape and scale parameters: $a = 4$ and $b = 0.25$. (We have also conducted additional numerical experiments under different parameters of the Gamma distribution, and under different distributions such as uniform and normal distributions, and found that our main insights remain directionally true.) The base parameters, as listed in Table 2.5, are loosely based on the U.S. influenza vaccine market and justified in the appendices. We vary these parameter values in our subsequent analysis for robustness of our results.

Table 2.5: Base Parameters Used in Numerical Experiments

c_r	c_e	w	p	α	β	γ
\$3	\$3	\$12	\$18	0.8	0.95	0.5

2.7.1 Performance of QF Contract

We evaluate the performance of the QF contract under different margins of the supply chain. Interestingly, the left panel of Figure 2.2 gives the following observation:

Observation 2.1 *The performance of the QF contract is near-optimal when the profit margin is either very high or very low, whereas the performance suffers when the profit margin is medium.*

The intuition is that when the profit margin is high (i.e., the unit cost is low), the optimal QF contract provides a generous return allowance and hence motivates the retailer to place a large order (see the right panel of Figure 2.2), which, in turn, motivates the manufacturer to choose a high early production quantity. (Note that although the supply chain efficiency is close to 100%, the supply chain is not perfectly coordinated when $Q_e^{FB} > 0$ according to Proposition 2.4.) When the profit margin of the supply chain is low, however, from our discussion in §2.3 we know that $Q_e^{FB} = 0$, meaning that the early production mode is not a

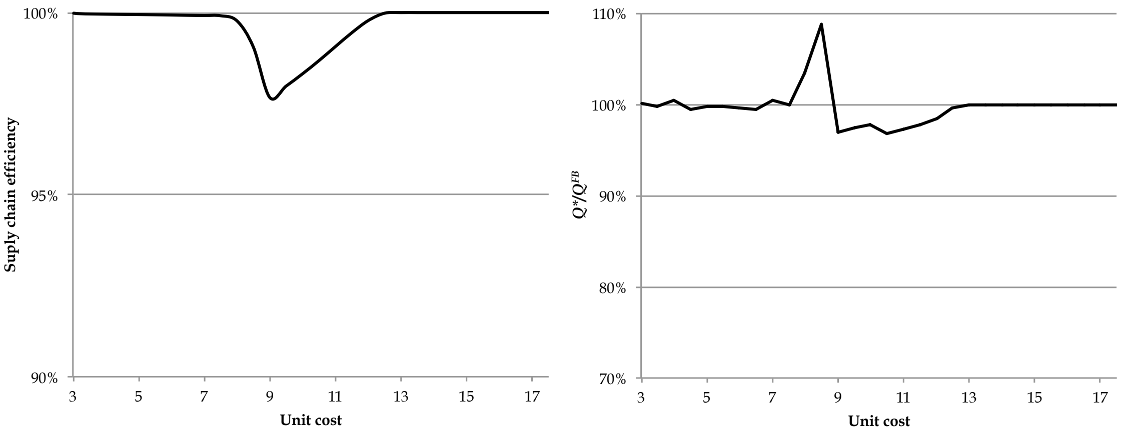


Figure 2.2: The performance of the QF contract under different production costs.

Note: The left panel shows the supply chain efficiency, and the right panel shows the ratio of the retailer's equilibrium ordering quantity (Q^*) to the first-best ordering quantity (Q^{FB}). We vary the production cost $c_r = c_e$ from \$3 to \$17.5, and set the wholesale price $w = 0.6p + 0.4c_r$. Other parameters follow the base parameters as listed in Table 2.5.

viable option. The supply chain can therefore be perfectly coordinated using a QF contract, as is the case where there is a single production mode.

Next, we examine the sensitivity of the performance of the QF contract with respect to the flexibility parameter (κ). As shown by the solid line in Figure 2.3, the QF contract exhibits robust and near-optimal performance for a wide range of the flexibility parameter around the optimal $\kappa^* = 0.60$. Figure 2.3 also shows the firms' profit shares under the QF contract. One immediate implication is that the manufacturer can choose a low κ to gain a higher profit share without significantly jeopardizing the efficiency of the supply chain.

2.7.2 Performance of LR Contract

Earlier in §2.6.2, Proposition 2.5 has shown that the LR contract coordinates the supply chain only when $w = c_e/\beta$, at which the retailer takes all of the supply chain profit. We now numerically examine the impact of different wholesale prices. Figure 2.4 shows that the LR contract can achieve satisfactory performance when w is sufficiently close to c_e/β (i.e., the retailer has a dominating bargaining power). For example, when $w = 1.5c_e/\beta$, the retailer's profit share is 83.6%, and the supply chain efficiency is 98.9%. However, the performance of the supply chain declines as w deviates further from c_e/β. For example,

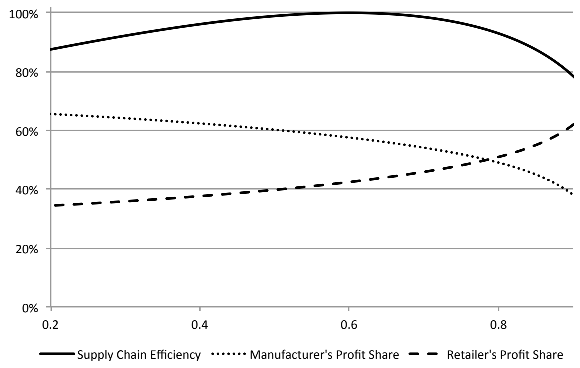


Figure 2.3: The supply chain efficiency and firms' profit shares under different values of κ in the QF contract.

Note: All the parameters follow the base parameters as listed in Table 2.5.

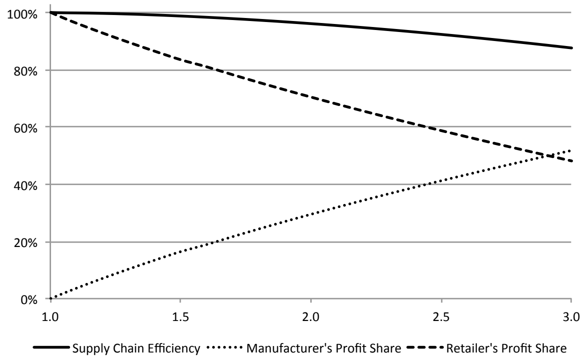


Figure 2.4: The supply chain efficiency and firms' profit shares under different wholesale prices.

Note: The x-axis is the ratio of w to c_e/β. Other parameters follow the base parameters as listed in Table 2.5.

when $w = 2.2c_e/\beta$, the retailer's profit share and the supply chain efficiency drop to 65.6% and 94.8%, respectively.

2.7.3 Evaluation of Contracts Used in Practice

We now evaluate the performance of the sample contracts used by Manufacturers A and B during the period 2009–2011, as shown in Table 2.1. For both firms, we assume that $c_e = c_r = \$3$, and $p = \$18$. We use the actual wholesale prices for 0.5ml syringe offered by Manufacturers A and B, respectively, during the 2009-2010 season. (For confidentiality, we do not report the actual prices here. The two wholesale prices are roughly between \$8

Table 2.6: Performance of the Supply Contracts Used by Manufacturers A and B

{α, β, γ}	D-QF Used by A (2008-2011)			LR Used by B (2009-2010)			Wholesale Price Contract Used by B (2008-2009; 2010-2011)		
	M Profit Share %	R Profit Share %	SC Efficiency %	M Profit Share %	R Profit Share %	SC Efficiency %	M Profit Share %	R Profit Share %	SC Efficiency %
{0.6,0.9,0.1}	42.49	57.51	97.46	45.52	54.48	90.09	41.08	58.92	91.35
{0.6,0.9,0.3}	42.94	57.06	98.25	45.60	54.40	90.53	41.48	58.52	83.30
{0.6,0.9,0.5}	44.03	55.97	98.75	45.84	54.16	90.47	42.82	57.18	72.50
{0.6,0.95,0.1}	43.50	56.50	97.79	46.34	53.66	90.20	41.08	58.92	90.14
{0.6,0.95,0.3}	44.04	55.96	98.23	46.37	53.63	90.30	41.48	58.52	81.44
{0.6,0.95,0.5}	45.18	54.82	98.69	46.50	53.50	90.23	42.82	57.18	70.38
{0.6,0.99,0.1}	44.29	55.71	97.77	46.97	53.03	90.02	40.89	59.11	92.56
{0.6,0.99,0.3}	44.86	55.14	98.07	46.98	53.02	90.02	40.89	59.11	92.56
{0.6,0.99,0.5}	46.07	53.93	98.60	47.00	53.00	90.00	40.92	59.08	92.54
{0.7,0.9,0.1}	42.37	57.63	97.00	45.43	54.57	90.24	41.07	58.93	92.02
{0.7,0.9,0.3}	42.69	57.31	97.87	45.79	54.21	84.34	41.45	58.55	86.10
{0.7,0.9,0.5}	43.49	56.51	98.37	47.04	52.96	76.03	42.72	57.28	77.93
{0.7,0.95,0.1}	43.33	56.67	97.48	46.36	53.64	90.13	41.07	58.93	91.00
{0.7,0.95,0.3}	43.72	56.28	97.90	46.39	53.61	90.28	41.45	58.55	84.38
{0.7,0.95,0.5}	44.52	55.48	98.28	46.49	53.51	90.25	42.72	57.28	75.92
{0.7,0.99,0.1}	44.09	55.92	97.52	46.98	53.02	90.02	41.07	58.93	89.88
{0.7,0.99,0.3}	44.50	55.50	97.76	46.98	53.02	90.02	41.45	58.55	82.93
{0.7,0.99,0.5}	45.34	54.66	98.17	47.00	53.00	90.00	42.72	57.28	74.32
{0.8,0.9,0.1}	42.64	57.36	97.69	46.00	54.00	90.41	41.07	58.93	92.45
{0.8,0.9,0.3}	42.46	57.54	97.42	46.29	53.71	86.90	41.36	58.64	88.96
{0.8,0.9,0.5}	42.95	57.05	97.87	47.30	52.70	81.42	42.37	57.63	83.61
{0.8,0.95,0.1}	43.18	56.82	97.12	46.39	53.61	90.01	41.07	58.93	91.82
{0.8,0.95,0.3}	43.42	56.58	97.54	46.41	53.59	90.24	41.36	58.64	87.45
{0.8,0.95,0.5}	43.92	56.08	97.84	46.48	53.53	90.26	42.37	57.63	81.78
{0.8,0.99,0.1}	43.89	56.11	97.27	46.98	53.02	90.01	41.07	58.93	90.83
{0.8,0.99,0.3}	44.16	55.84	97.43	46.99	53.01	90.02	41.36	58.64	86.10
{0.8,0.99,0.5}	44.68	55.32	97.71	47.00	53.00	90.01	42.37	57.63	80.28
Average	43.80	56.20	97.85	46.50	53.50	88.98	41.61	58.39	85.35

Note: "M" stands for the manufacturer, "R" for the retailer, and "SC" for the supply chain.

and $10, and differ only slightly.) When Manufacturer B used the wholesale price contract during the 2008-2009 and 2010-2011 seasons, its wholesale price was lower than that of Manufacturer A by 15% on average. Thus, for fair comparison between the LR contract and the wholesale price contract, we use the wholesale price for the wholesale price contract that is lower by 15% than that for the LR contract. Table 2.6 shows the firms' profit shares and the supply chain efficiency under different combinations of (α, β, γ). We observe that the supply chain's efficiency is above 80% in most of the instances, suggesting that both firms' contracts can achieve reasonable performance. Furthermore, we make the following recommendations to improve these supply contracts:

(i) Manufacturer A can keep the current D-QF contract but consider using higher return allowance proportions to increase the supply chain efficiency: our additional numerical study suggests that in the optimal D-QF contracts, on average, $\kappa_1^* = 47\%, \kappa_2^* = 58\%$, with a supply chain efficiency of 99.8%. On the other hand, we find that if Manufacturer A is only concerned about maximizing its own profit, then it is optimal to set $\kappa_1^* = 0$ and $\kappa_2^* = 95\%$, which is close to that of Manufacturer C (see §1).

(ii) For the ease of implementation, Manufacturer A can switch to the QF contract with κ chosen around 50%, at which point the supply chain efficiency is 99.86%. Our further numerical study shows that if Manufacturer A is only concerned about its own profit maximization, then it is optimal to use a lower κ (on average, $\kappa^* = 9\%$, with a supply chain efficiency of 91.33%).

(iii) Manufacturer B should switch from its current wholesale price contract back to the LR contract. Doing this will improve the supply chain efficiency (from the average of 85.35% to 88.98% as shown in Table 2.6) as well as the manufacturer's profit share (from the average of 41.61% to 46.50%). For further improvement, we recommend that Manufacturer B use the BLR contract for it gives the supply chain a higher level of efficiency and flexibility of profit division.

2.8 Extensions

We now discuss two extensions to our model. §2.8.1 extends the distribution of early production mode to be continuous. §2.8.2 discusses the case when the performance metric is social welfare instead of the efficiency of the supply chain.

2.8.1 When Early Production Outcome Is Continuous

So far we have captured the uncertainty in product design by assuming that early production will be completely wasted if the projected design does not match the final design. While this is true in the influenza vaccine industry, a manufacturer in other industries might be able to recover a proportion of the products. This is particularly true in the apparel and footwear industry.

We now assume that the early production results in a continuous, uncertain outcome. Specifically, a proportion $\tilde{\theta}$—a random variable with a pdf of $h(\cdot)$, a cdf of $H(\cdot)$, and a mean of θ—of the units in the early production mode is retained after the finalized design is announced. The sequence of events remains the same as in Figure 2.1, but the relationship between Q_e and Q_r is now generalized as $Q_r = Q - \tilde{\theta} Q_e$. The manufacturer's expected production cost is now $E[c_e Q_e + c_r Q_r] = E[c_e Q_e + c_r(Q - \tilde{\theta} Q_e)] = c_r Q + (c_e - \theta c_r) Q_e$. The expected selling quantity, given $\tilde{\theta}$, can be represented as

$$Z(Q, Q_e|\tilde{\theta}) = \begin{cases} \min\{Q, d\} & \text{with probability } \alpha \\ \min\{\tilde{\theta} Q_e, d\} + \min\{Q - \tilde{\theta} Q_e, (1-\gamma)(d - \tilde{\theta} Q_e)^+\} & \text{with probability } 1-\alpha. \end{cases}$$

Despite the more general representations of the production cost and the expected sales quantity, we can extend the BLR contract to the case in which early production yield a continuous outcome set. The following corollary is immediate from (2.3) by replacing β with θ:

Corollary 2.2 *When the early production yields a continuous outcome set, a BLR contract with* $b^*_{BLR} = \frac{p(\theta w - c_e)}{\theta p - c_e}$ *and* $\rho^*_{BLR} = \frac{(p-w)(c_e - \theta c_r)}{(1-\alpha) w (\theta p - c_e)}$ *coordinates the supply chain.*

We can also show that the LR contract coordinates the supply chain only when $w = c_e/\theta$ and the retailer takes all the profit of the supply chain, and the supply chain efficiency deteriorates when w deviates from c_e/θ. Numerically, we have also verified that the QF contract never coordinates the supply chain unless the early production is not viable, while the D-QF contract can coordinate the supply chain in some cases. Therefore, the qualitative results will remain unchanged when early production uncertainty follows a continuous distribution.

2.8.2 Social Welfare

We have focused our attention on designing supply contracts that maximize the supply chain's expected profit. Now we bring consumers' benefits into context and analyze the impact of supply contract design on social welfare. Consistent with extant literature, we define the social welfare as the sum of consumers' surplus and the supply chain's profit. We assume that consumers' aggregated benefits from consuming the product is $V(z)$ when the selling quantity is $Z(Q, Q_e) = z$, where $V(\cdot)$ is assumed to be a concave increasing function, reflecting the diminishing marginal improvement in social welfare from a higher coverage. The expected social welfare $W(Q, Q_e)$ can be represented similar to (2.4) as follows:

$$W(Q, Q_e) = E[V(Z(Q, Q_e))] - [(c_r + c_o)Q + (c_e - \beta c_r)Q_e]. \tag{2.13}$$

In the influenza vaccine industry, when the initial fraction of infected population is small, Chick et al. (2008) show that the net benefits from vaccination can be represented in a piecewise-linear form. We define v as each consumer's surplus from vaccination, and rewrite (2.13) as

$$W(Q, Q_e) = vE[Z(Q, Q_e)] - [c_r Q + (c_e - \beta c_r)Q_e]. \tag{2.14}$$

Comparing (2.14) with (2.4), we see that the expected social welfare differs from the expected supply chain profit mainly in the value associated with the expected selling quantity $Z(Q, Q_e)$. When $v > p$, the social planner values each unit of sold product more than the supply chain would, so we expect that the selling quantity in the social optimum is higher than that in the market equilibrium.

One way to address the social welfare gap is to introduce a third party that provides a subsidy s to the retailer—rather than the manufacturer—for each unit of products that are shipped and sold to consumers. At $s = v - p$, the retailer's total unit revenue from a consumer and from the third party is exactly v, and the analysis of all the supply contracts remains the same. We therefore expect that using government subsidies, together with well-calibrated supply contracts between the manufacturer and the retailer, would bring the vaccine market closer to the social optimum.[7]

[7] While Chick et al. (2008, 2012) consider the government as a central purchaser, we consider the contract between a manufacturer and a *single* retailer (among numerous retailers), and hence the demand of the retailer does not represent the demand of the entire population. Therefore, it is not practically possible to estimate the value of vaccination to a consumer of that retailer. Our discussion in this section intends to illustrate the need of a government subsidy to a retailer, but not to propose a method to estimate the right amount of the subsidy, which is left for future research.

2.9 Concluding Remarks

In this paper, we study a contract design problem for the U.S. influenza vaccine supply chain that faces uncertainties in the design, delivery, and demand of the product. To mitigate the risk of late delivery, the manufacturer operates in two production modes: the regular production mode that starts production after the design uncertainty is resolved, and the early production mode that starts production before the design is finalized. The manufacturer faces a trade-off between the informational advantage of the regular production, and the delivery advantage of the early production. Currently, the industry is experimenting different contract terms aiming to improve the efficiency of the supply chain. This makes our study both explanatory and prescriptive.

Our analysis reveals that without the careful design of a supply contract, a vicious incentive cycle may arise: because the manufacturer bears the risk associated with the early production mode, it lacks incentive to improve the on-time delivery, which reduces the retailers order size in anticipation of truncated demand; and this further discourages the manufacturer from making efforts to improve its delivery performance. Surprisingly, we have found that the revenue sharing contract may lead to zero delivery-improving effort, and perform even worse than the wholesale price contract. Therefore, we propose two coordinating contracts: the Delivery-time-dependent Quantity Flexibility (D-QF) contract, and the Buyback-and-Late-Rebate (BLR) contract. In view of the complexity of these contracts, we move on to consider their simpler counterparts that are easier to implement. We conduct extensive numerical experiments based on realistic values from the U.S. influenza vaccine industry, and show that the QF contract, despite its low informational requirement and simple contract structure, can achieve near-optimal performance when the gross margin of the product is either very high or very low. A simple LR contract, by contrast, performs well only when the wholesale price is relatively low such that the retailer takes a major portion of the supply chain profit. These insights can help practitioners design supply contracts for improving on-time delivery performance of the influenza vaccine supply chain, and potentially those of other supply chains with similar characteristics (e.g., the apparel and footwear industry).

There are several interesting future research avenues. Motivated also by the influenza vaccine industry, Federgruen and Yang (2008) and Cho and Tang (2012) consider supply chain models under uncertain demand and supply—the former studies an issue of selecting a portfolio of suppliers under a fixed wholesale price, and the latter studies the benefits of dynamic ordering and selling under a wholesale price contract. Studying sophisticated contracts among multiple suppliers in the dynamic setting will be an interesting direction for future research. Also, it would be interesting to study the manufacturers' contract choices in a competitive setting, i.e., how one manufacturer's choice of a certain contract type may affect its competitors' decisions.

Chapter 3

The Welfare Consequences of the Donor Priority Rule

> We assume some of the most peculiar and temporary of our late advantages as natural, permanent, and to be depended on, and we lay our plans accordingly... On this sandy and false foundation we scheme for social improvement and dress our political platforms...
>
> John Maynard Keynes: *The Economic Consequences of The Peace* (1919)

3.1 Background

The U.S. is experiencing an ongoing organ shortage crisis, with about 18 people dying while waiting for transplant each day, and a new candidate added to the waiting list every 10 minutes (Organdonor.gov 2012). Furthermore, the waiting list of transplant candidates outgrows the registry of potential organ donors: between 1989 and 2009, the number of people wait-listed for an organ increased 4.89 times, but the number of cadaveric organ donors (i.e., those who agree to donate their organs in case of premature brain death) grew merely 1.47 times. As cadaveric organs remain to be the primary source of organs for transplantation, one major cause of the current organ crisis is the low share of registered organ donors. Currently only 40% of the U.S. population are registered organ donors, although the public has an 85% approval rate of organ donation (Gallup 2005).

A myriad of initiatives have been proposed to encourage more people to add their names to the organ donor registry. The contemporary public discourse mostly focuses on education and promotion to enhance the public's awareness of the benefits of organ donation. On April 2, 2012, President Barack Obama proclaimed each April afterwards as the National Donate Life Month to "call upon health care professionals, volunteers, educators, government agencies, faith-based and community groups, and private organizations to join forces to boost the number of organ and tissue donors." On May 1, 2012, Facebook announced a new sharing function that enabled its users to advertise their donor status on their timelines, in the hope that this move would exert peer pressure on people who have not registered as organ donors.

Health economists, on the other hand, have long been at the forefront of proposing to provide monetary incentives (in cash or non-cash forms) to individuals for registering as potential cadaveric organ donors. While few would contest that doing so would lead to an immediate jump in the number of organ donors, this market-based approach been widely criticized as it allegedly "fosters class distinctions (and exploitation), infringes on the inalienable values of life and liberty, and is therefore ethically unacceptable" (Delmonico et al. 2002).

In addition to enhancing public awareness and providing monetary incentives, there are two popular initiatives that have been weighed by federal and state governments as well as non-government organizations: (i) Presumed consent (aka "opt-out") policy, which, in contrary to the current practice in the U.S., automatically registers people as potential organ donors (e.g., when they apply for a driver license) unless they follow required procedures to opt out of the organ donation program. The legislation of presumed consent has been endorsed by various studies (e.g., Abadie and Gay 2006) but faces many hurdles, including the public's fear of misrepresentation of individuals' willingness to donate. (ii) Donor priority rule, which provides a priority status to individuals registering to become potential organ donors. Under the rule, in the event that registered donors need organ transplants, they are given higher priority in receiving cadaveric organs than non-donors. Our paper focuses on analyzing the donor priority rule. We develop a theoretic model of the system of donor registration and organ allocation. As we focus on the broad implications of adopting the rule, we do not restrict ourselves to a specific type of organ (e.g., kidney, liver, tissue).

In modeling the tradeoffs behind each individual's decision to register as a potential organ donor, we follow Kessler and Roth (2012) to assume that each individual has a cost of donating. Different from Kessler and Roth (2012), however, we capture each individual's utility from organ transplantation using the expected total quality-adjusted life expectancy (QALE) by applying the approximation results from the queueing literature (e.g., Zenios 1999). Our analysis shows that when all individuals are homogeneous in their health status, the introduction of the donor priority rule will expand the size of the donor registry, increase the overall availability of obtaining an organ, and result in increased social welfare. This result is consistent with the findings by Kessler and Roth (2012). When the individuals are heterogeneous, however, we show that, in contrast to what Kessler and Roth (2012) predict, the introduction of the donor priority rule can indeed *reduce* social welfare. This is because under the rule, even for individuals with the same cost of donating, as they can have different probabilities of needing organ transplants in the future, they will respond differently to the rule in their decisions as to whether to register to be a potential organ donor or not. Specifically, we show that ceteris paribus, the donor priority rule provides higher incentives to high-risk individuals than to low-risk individuals, and essentially leads to a pool of donated organs with an average quality lower than that of the general population. When this disproportionality of incentives becomes significant enough, we show that the resultant social welfare loss can outweigh the social welfare gain from the expanded organ donor registry.

We proceed to show that a simple freeze period remedy, under which an individual may not enjoy a higher queueing priority until a period of time has passed since his registration as an organ donor, can overcome the aforementioned quality distorting effect. When the freeze period remedy is implemented in conjunction with the donor priority rule, the average quality of the donated organs can be restored to the population average level. The underlying reason for this improvement is that the freeze period remedy essentially provides a disincentive for individuals to become organ donors, but the level of disincentive differs for individuals with different risk levels such that high-risk individuals are more discouraged than low-risk ones. We prove that the remedy, if designed properly, can (i) help fix the biased incentive structure due to the donor priority rule, and (ii) expand the donor base compared to the case without the donor priority rule. Thus the freeze period remedy, in conjunction with the donor priority rule, will lead to a better social outcome compared to the current organ donation system.

The rest of the paper is organized as follows. Section 3.2 reviews the relevant literature. In §3.3, we describe our analytical framework. In §3.4, we consider the case where all the individuals have homogeneous health status. Section 3.5 models population heterogeneity in health status. Section 3.6 proposes a freeze period remedy. Section 3.7 concludes the paper with a summary of our key insights and future research directions.

3.2 Literature

We add to a thin but growing body of Operations Management (OM) literature on organ transplantation services, most of which focuses on organ allocation (e.g., Su and Zenios 2004, 2006; Akan et al. 2012; Ata et al. 2012) and surgical decisions (e.g., Howard 2002; Alagoz et al. 2004). To the best of our knowledge, we are the first to develop an OM model to examine the organ donation policy. Here we briefly review several papers on organ allocation that are most relevant to our work. Su and Zenios (2004) develop and analyze a queueing model to examine the role of patient choice in the kidney transplant waiting system, and highlight the conflict between equity and efficiency in kidney allocation. They show that under the last-come first-served priority discipline, the competitive equilibrium is socially optimal. Our paper, with a specific focus on organ donation, also considers a priority queueing discipline, but the priority is tied to each individual's organ donor status rather than the sequence of arrivals. Su and Zenios (2006) propose an organ allocation method where heterogeneous patients have to declare which types of kidneys they would be willing to accept at the time they join the waiting list (rather than at the time they are offered the kidney). By doing so, a lengthy search at time of transportation is eliminated, and each candidate's private information is reflected in the allocation process. Our paper highlights the impact of individual heterogeneity as well and shows that individuals' heterogeneous health status can influence their organ donation behavior and, in turn, the average quality of the donated organs.

In the vast health economics literature, there is a paucity of rigorous treatment of organ donation through modeling the donor priority rule. The only paper that analytically examines the same issue is by Kessler and Roth (2012), who address the impact of the donor priority rule mainly through behavioral experiments. Their analytic model, built to illustrate their experimental findings, only considers how the donor priority rule affects an individual's probability of receiving an organ. Our paper, by contrast, builds a queueing model of the organ allocation process, and uses an individual's QALE, rather than the probability of receiving organs, as the measure of an individual's utility. Our setup gives rich and interesting insights into the social welfare consequences of the donor priority rule.

3.3 Modeling Framework

Each individual can be in one of three states: healthy (in a condition not requiring an organ transplant), sick and in need of an organ transplant, or dead from brain death. Two types of events are central to our modeling of the organ allocation system: some healthy individuals become sick and are in need of an organ, while other healthy individuals suffer from brain death and can potentially provide a number of organs. We model these events as two separate, independent stochastic processes. We denote by θ the arrival rate of individuals in need of organs, and ϕ the rate of all brain deaths; both processes are assumed to be Poisson processes without loss of generality. Letting n denote the number of organs each potential donor can provide, the maximum possible arrival rate of organs is ϕn. We assume that $\phi n > \theta$, meaning that the organs from all the brain deaths—whether organ donor registry or not—are adequate to support those in need of organs.[1]

Each individual incurs a cost to register as an organ donor. The cost is denoted by c and has a support of $(-\infty, \infty)$ and a cumulative density function $F(\cdot)$. A positive cost c corresponds to the case where the individual is faced with a burden to overcome to reg-

[1] This is a realistic assumption given the statistics about number of brain deaths and listed patients. In 2007, for example, there were 2.5 million premature brain deaths in the U.S., which is more than twenty times the number of patients on the waiting list for cadaveric organs (Organdonor.gov 2012).

ister to be a potential organ donor. For example, there have been documented fears that physicians might not try their best to save registered organ donors' lives (Teresi 2012). As another example, certain religious beliefs disfavor the practice of organ donation (Bruzzone 2008). When an individual has a cost $c < 0$, it means that the individual earns a positive non-monetary gain (e.g., social recognition, self-fulfilment) from registering to be an organ donor (Prottas 1983).

Following the organ transplantation literature (e.g., Su and Zenios 2006), we use QALE to measure the utility of an individual who needs an organ transplant. An individual's QALE is written as

$$u = \alpha D + \beta \pi T,$$

where α is the quality-of-life score while on the waiting list; D is the individual's life expectancy from the time he is put on the waiting list to the time he dies or receives an organ, whichever comes earlier; β is the quality-of-life score after transplantation; π is the probability of receiving an organ; and T is the individual's post-transplantation life expectancy.

For a healthy individual, the tradeoff behind the decision of becoming organ donor involves the cost of donating versus the potential benefits from organ donation (if any). It is evident that ceteris paribus, an individual with a higher cost will have a lower incentive to become an organ donor. In other words, given any organ donation policy, there must exist a threshold $x \in (-\infty, \infty)$ such that all the individuals with $c \leq x$ will become organ donors, and all the individuals with $c > x$ will not. Given x, the corresponding donation rate (i.e., the proportion of the population who are donors) is $F(x)$. As $F(x)$ increases in x, a larger x means that there is a higher share of organ donors. We define a constant $\hat{x}$ at which $F(\hat{x})\phi n = \theta$, that is, $\hat{x} = F^{-1}(\theta/(\phi n))$. In other words, $F(\hat{x})$ is the share of registered organ donors at which the supply rate of organs is equal to the demand rate.

Let d denote a sick individual's life expectancy without organ transplantation. Under a cost threshold of x, we can approximate the individual's pre-transplantation life expectancy as

$$D(x) = d\left(\frac{\theta - F(x)\phi n}{\theta}\right)^{+} = \begin{cases} \frac{\theta - F(x)\phi n}{\theta} \cdot d & \text{if } x \leq \hat{x} \\ 0 & \text{otherwise.} \end{cases} \tag{3.1}$$

The probability for each individual to receive an organ π can be approximated as

$$\pi(x) = \min\left\{1, \frac{F(x)\phi n}{\theta}\right\} = \begin{cases} \frac{F(x)\phi n}{\theta} & \text{if } x \leq \hat{x} \\ 1 & \text{otherwise.} \end{cases} \tag{3.2}$$

The above approximations were developed in Zenios (1999).

3.3.1 Benchmark: Social Optimum

Before proceeding to investigate the effect of a specific policy, we first characterize the social optimum. Since all individuals are homogeneous in their ex ante expected utility of receiving an organ, in the social optimum, they have the same probability of receiving an organ. In the social optimum, the donors, as chosen by the social planner, are the individuals with low costs of donating. Therefore, the social optimum is dictated by a threshold, denoted by x^{SO}, such that all individuals with costs lower than x^{SO} will register to become donors, and those with costs higher than x^{SO} will not. The social welfare as a function of x can thus be written as the expected aggregate QALE of all the listed individuals less the aggregated costs of donating of all the registered organ donors (i.e., those whose cost of donating is lower than x):

$$\begin{aligned} W_s(x) &= \theta(\alpha D(x) + \beta\pi(x)T) - \mathbb{E}[(c|c \leq x)]F(x) \\ &= \begin{cases} \alpha d\left(\theta - F(x)\phi n\right) + \beta T F(x)\phi n - \int_{-\infty}^{x} c f(c)\,dc & \text{if } x < \hat{x}, \\ \theta\beta T - \int_{-\infty}^{x} c f(c)\,dc & \text{otherwise.} \end{cases} \end{aligned} \tag{3.3}$$

Lemma 3.1 *The socially efficient cost threshold is* $x^{SO} = \min\left\{\left(\beta T - \alpha d\right)\phi n, \hat{x}\right\}$.

Proof. When $x < \hat{x}$, we have from (3.3) that

$$dW_s(x)/dx = ((\beta T - \alpha d)\phi n - x)f(x).$$

This implies the regional maximum for $x \in (-\infty, \hat{x})$ is achieved at $\left(\beta T - \alpha d\right)\phi n$ or $\hat{x}$, whichever is lower. If $x > \hat{x}$, we observe from (3.3) that $W_s(x)$ decreases in x, indicating that the regional maximum for $x \in (\hat{x}, \infty)$ is achieved at $\hat{x}$. Taken together, the socially optimal threshold x^{SO} is $\left(\beta T - \alpha d\right)\phi n$ or $\hat{x}$, whichever is lower. *Q.E.D.*

3.4 Preliminary: Homogeneous Health Status

As a benchmark, we consider in this section the case where individuals have heterogeneous costs of donating but homogeneous health status. We first characterize the equilibrium without the donor priority rule. Then we characterize the equilibrium after the introduction of the donor priority rule. We then compare the social welfare and show that introducing the donor priority rule leads to increased social welfare.

3.4.1 Without Priority Rule

As a benchmark, we consider the case where the allocation of organs follows a first-come-first-serve (FCFS) queueing discipline, and registered organ donors do not enjoy any priority over non-donors if they are in need of organs in the future. Because each individual gains no benefits from becoming an organ donor, only those with $c < 0$ have the incentive to do so. Therefore, in equilibrium, the threshold cost, denoted by x^*_{np}, is zero; the share of organ donors (i.e., the proportion of the population registering as potential donors) is $F(0)$. The resultant arrival rate for the supply process of organs is therefore $n\phi F(0)$. We assume that $n\phi F(0) < \theta$, meaning that the supply rate of organs is lower than the demand rate, which is in line with the current situation of the U.S. organ transplantation system (Organdonor.gov 2012).

Each sick individual's pre-transplantation life-expectancy, by (3.1), is $D(x^*_{np}) = d(\theta - F(0)\phi n)/\theta$; the probability for each sick individual to receive an organ, by (3.2), is $\pi(x^*_{np}) =$

$(F(0)\phi n)/\theta$. The social welfare with no donor priority rule, denoted by W_{np}, is the expected aggregate QALE of all the sick individuals less the costs of donating of all those who choose to register as organ donors. That is,

$$\begin{aligned} W_{np} &= \theta(\alpha D(x_{np}^*) + \beta\pi(x_{np}^*)T) - \mathbb{E}[(c|c \leq 0)]F(0) \\ &= \alpha d\left(\theta - F(0)\,\phi n\right) + \beta F(0)\,\phi n T - \mathbb{E}[(c|c \leq 0)]F(0). \end{aligned} \tag{3.4}$$

Comparison with Social Optimum

In the social optimum characterized in §3.3.1, we have $x^{SO} = \min\left\{\left(\theta\beta T - \alpha d\right)\phi n, \hat{x}\right\}$. Since $n\phi F(0) < \theta$, we have $\hat{x} = F^{-1}(\theta/(\phi n)) > 0$. In addition, the condition $\beta T > \alpha d$ gives $\left(\beta T - \alpha d\right)\phi n > 0 = x_{np}^*$. Therefore, $x^{SO} = \min\{(\beta T - \alpha d)\phi n, \hat{x}\} > 0$. In other words, without the priority rule, the donation rate is strictly below the socially optimal level.

3.4.2 With Priority Rule

Now we consider the case where the registered donors are given priority over non-donors to receive organs when they need cadaveric organs. Due to the introduction of the donor priority rule, there are two queues of patients waiting for organs: a priority queue consisting of registered organ donors, and a regular queue consisting of non-donors. When an individual decides to be an organ donor, she is essentially purchasing an option to join a priority queue in the future. Therefore, the individual would benefit from a larger organ pool provided by more registered organ donors. However, as more individuals become potential donors, the relative attractiveness of organ donation declines, as the surplus supply rate for the non-donors increases.

To seek the point at which the equilibrium is reached, we first need to derive an organ donor's utility. We use x_p to denote the cut-off cost at which an individual is indifferent as to whether to elect to be a donor or not. Given x_p, the supply rate of organs is now $\phi n F(x_p)$, and the total demand rate is θ, of which $F(x_p)\theta$ is the arrival rate of donors, and the remaining $(1 - F(x_p))\theta$ is the arrival rate of non-donors. There are two waiting lists—donor list and non-donor list—for organs. Because a donor has a higher priority than a non-donor, they have different utility values. Below we determine their utility values under an given x_p.

As $\theta < \phi n$, by (3.1) and (3.2), a registered donor's expected waiting time for an organ is zero, and probability of receiving a transplant is one. Therefore, a donor's net utility is his QALE in the event that he will need a organ less his cost, that is, $u_d = \beta\theta T - c$.

A non-donor has lower priority than the donor, and faces scarce supply. The total arrival rate of sick non-donors is $\lambda_n = (1 - F(x_p))\theta$, and the supply rate of organs for non-donors is $\mu_n = F(x_p)(\phi n - \theta)$. Each non-donor's pre-transplantation life expectancy and probability of receiving an organ can be computed using (3.1)–(3.2):

$$D_n(x_p) = \frac{\lambda_n - \mu_n}{\lambda_n} \cdot d = \frac{\theta - \phi n F(x_p)}{(1 - F(x_p))\theta} \cdot d, \text{ and}$$
$$\pi_n(x_p) = \frac{\mu_n}{\lambda_n} = \frac{F(x_p)(\phi n - \theta)}{(1 - F(x_p))\theta}.$$

The following proposition characterizes the equilibrium under the donor priority rule.

Proposition 3.1 *Under the donor priority rule and a population with homogeneous health status, in equilibrium, only those individuals with costs below a cutoff cost x_p^* will elect to register to become organ donors, where x_p^* satisfies*

$$x_p^* = \frac{\theta - \phi n F(x_p^*)}{1 - F(x_p^*)} \cdot (\beta T - \alpha d). \tag{3.5}$$

Proof. An individual with the cutoff cost x_p^* is indifferent as to whether to register to become an organ donor or not, that is, $\theta\beta T - x_p^* = \theta(\alpha D(x_p^*) + \beta T \pi_n(x_p^*))$, which can be reorganized as (3.5). *Q.E.D.*

Based on Proposition 3.1, we can write the social welfare under the donor priority rule as

$$W_p = \underbrace{\int_{-\infty}^{x_p^*} \left(\theta\beta T - c\right) f_c dc}_{\text{Donors' welfare}} + \underbrace{\int_{x_p^*}^{\infty} \theta \left(\alpha d \cdot \frac{\lambda_n - \mu_n}{\lambda_n} + \beta T \cdot \frac{\mu_n}{\lambda_n}\right) f(c) dc}_{\text{Non-donors' welfare}}$$
$$= \theta\beta T - E[(c | c \leq x_p^*)] F\left(x_p^*\right) - x_p^* \left(1 - F\left(x_p^*\right)\right). \tag{3.6}$$

Proposition 3.1 gives the following corollaries:

Corollary 3.1 $x_p^* > x_{np}^* = 0.$

Corollary 3.2 x_p^* *increases in* θ*, and decreases in* ϕ*.*

Proof. We have from (3.5) that

$$1 - \frac{\theta}{\phi n} = \left(1 - \frac{x_p^*}{\phi n(\beta T - \alpha d)}\right)\left(1 - F(x_p^*)\right). \tag{3.7}$$

As θ increases, the left-hand side of (3.7) decreases, requiring a higher x_p^* to decreases the right-hand side of (3.7) and balance the equation. Similarly, we can show that x_p^* decreases in ϕ. *Q.E.D.*

The above corollary suggests that more individuals will register to become potential organ donors if the probability of becoming sick and needing an organ increases, and vice versa. In addition, fewer individuals will register to become potential organ donors when brain deaths occur more frequently, providing a higher supply rate of organs.

Comparison with Social Optimum. Corollary 3.1 implies that introducing the donor priority rule will lead to a higher donation rate. However, we can verify that the equilibrium donation rate is still below the social optimal donation rate, i.e. $x_p^* < x^{SO}$. To see this, recall that $x^{SO} = \min\left\{(\beta T - \alpha d)\phi n, \hat{x}\right\}$. On one hand, as $\theta - n\phi F(x_p^*) > 0$ (from (3.5)), we have $x_p^* < \hat{x} = F^{-1}\left(\theta/(\phi n)\right)$. On the other hand, we have

$$x_p^* = \frac{\frac{\theta}{\phi n} - F\left(x_p^*\right)}{1 - F\left(x_p^*\right)}\left(\beta T - \alpha d\right)\phi n < \left(\beta T - \alpha d\right)\phi n,$$

since $\theta < \phi n$. Hence we have $x_p^* < \min\{\hat{x}, \left(\beta T - \alpha d\right)\phi n\} = x^{SO}$.

3.4.3 Comparison of Social Welfare

We now compare the social welfare before and after the introduction of the donor priority rule. The result is summarized in the following proposition:

Proposition 3.2 *When the population is heterogeneous in cost of donating but homogeneous in health status, the introduction of the donor priority rule always increases the social welfare.*

Proof. We examine the difference in the social welfare before and after the introduction of the donor priority rule. By (3.4) and (3.6), we have

$$\begin{aligned} W_p - W_{np} &= \left(\theta - F(0)\,\phi n\right)\left(\beta T - \alpha d\right) + E\left(c|c \le 0\right)F(0) - E\left(c|c \le x_p^*\right)F\left(x_p^*\right) - x_p^*\left(1 - F\left(x_p^*\right)\right) \\ &= \left(\theta - F(0)\,\phi n\right)\left(\beta T - \alpha d\right) - \int_0^{x_p^*} cf(c)dc - x_p^*\left(1 - F\left(x_p^*\right)\right), \end{aligned}$$

which, by Proposition 3.1, can be rewritten as

$$\begin{aligned} W_p - W_{np} &= \left(\theta - F(0)\,\phi n\right)\frac{1 - F\left(x_p^*\right)}{\theta - F\left(x_p^*\right)\phi n}x_p^* - \int_c^{x_p^*} cf(c)dc - x_p^*\left(1 - F\left(x_p^*\right)\right) \\ &= x_p^*\left(F\left(x_p^*\right) - F(0)\right)\frac{1 - F\left(x_p^*\right)}{\theta/\left(\phi n\right) - F\left(x_p^*\right)} - \int_0^{x_p^*} cf(c)dc. \end{aligned} \tag{3.8}$$

Now, since $\theta < \phi n$, we have

$$\frac{1 - F\left(x_p^*\right)}{\theta/\left(\phi n\right) - F\left(x_p^*\right)} > 1,$$

which gives $W_p - W_{np} > x_p^*\left(F\left(x_p^*\right) - F(0)\right) - \int_0^{x_p^*} cf(c)dc > 0$. That is, the social welfare improves after the introduction of the donor priority rule. *Q.E.D.*

Giving donors priority to receive organs has two immediate effects: it increases the total costs of donating because a proportion of donors with positive costs are incentivized to register as donors, and it increases the supply of donors. Proposition 3.2 suggests that the second effect outweighs the first, leading to increased social welfare.

The following corollary further refines Proposition 3.2.

Corollary 3.3 *The social welfare difference $W_p - W_{np}$ increases in θ, and decreases in ϕ.*

Proof. The social welfare difference, by (3.8), can be rewritten as

$$W_p - W_{np} = x_p^*\left(F\left(x_p^*\right) - F(0)\right)\cdot\left(1 + \frac{1 - \theta/(\phi n)}{\theta/(\phi n) - F(x_p^*)}\right) - \int_0^{x_p^*} cf(c)dc,$$

which is increasing in x_p^*. As θ increases, we have from Corollary 3.2 that x_p^* increases, and so does the social welfare improvement. Similarly, we can show that the social welfare improvement decreases in ϕ. *Q.E.D.*

As θ increase or ϕ decreases, it becomes increasingly challenging for a candidate to be matched to an organ. The donor priority provides a stronger incentive for individuals to become registered donors and leads to a higher social welfare improvement.

3.5 The Unexpected Welfare Consequences of Donor Priority Rule

In the preceding section, we show that the introduction of the donor priority rule increases the social welfare. The baseline model allows individuals to be heterogeneous in their costs of donating. In this section, we extend our preliminary model to incorporate the second dimension of the heterogeneity among these individuals: their probability of becoming sick and needing an organ.

We categorize the whole population into two groups: high- and low-risk population. A high-risk individual is more likely to become sick and need an organ; in addition, a high-risk individual is more likely to have low-quality organs. To be specific, with probability p_H, an individual is high-risk; with probability p_L, an individual is low-risk. We have $p_H + p_L = 1$. We now break the demand process into two separate, independent subprocesses such that the arrival rate for high-risk individuals to become sick and need organs is θ_H, and the arrival rate for low-risk individuals to become sick and need organs is θ_L. The total arrival rate of individuals needing organ is now $p_H\theta_H + p_L\theta_L$. As the two population groups are different in their potential needs for organs due to their different health status, it is reasonable to assume that they also differ in the quality of their organs. The quality of an individual's organs is measured in terms of the post-transplantation life expectancy of another individual who receives one of these organs. We use $G_H(t)$ and $G_L(t)$ to denote the respective c.d.f.'s of the post-transplantation life expectancy associated with organs receiving from a high-risk and low-risk individual. We make the following assumptions:

Assumption 3.1 *The life expectancy of a person that receives an organ from a high-risk individual is lower than that from a low-risk individual, that is,* $\mathbb{E}[G_H(t)] = T_H < \mathbb{E}[G_L(t)] = T_L$.

Assumption 3.2 *The total demand rate for donated organs is lower than the maximum possible organ supply rate, that is, $p_H\theta_H + \pi_L\theta_L < \phi n$.*

Assumption 3.3 *The arrival rate for a high-risk individual to become sick and need an organ is higher than that for a low-risk individual, that is, $\theta_H > \theta_L$.*

3.5.1 Without Priority Rule

We first consider the case where the priority rule is not enforced. Similar to the preliminary model, only those with negative costs of donating have the incentive to register as organ donors. In other words, the cut-off cost in equilibrium is $x_i^* = 0$. Hence the aggregated post-transplantation life expectancy is

$$T_a = \sum_{i=L,H} p_i T_i.$$

The social welfare is thus

$$\begin{aligned} W_{np}^h &= \sum_{i=L,H} p_i\theta_i \left(\alpha d \frac{\sum_{i=L,H} p_i\theta_i - F(0)\,\phi n}{\sum_{i=L,H} p_i\theta_i} + \beta T_a \frac{F(0)\,\phi n}{\sum_{i=L,H} p_i\theta_i} \right) - \mathbb{E}[(c|c \le 0)]F(0) \\ &= \alpha d \left(\sum_{i=L,H} p_i\theta_i - F(0)\,\phi n \right) + F(0)\,\phi n \beta T_a - \mathbb{E}[(c|c \le 0)]F(0). \end{aligned}$$

3.5.2 With Priority Rule

We now introduce the priority rule to the organ allocation system. One key aspect different from our analysis in the preliminary model is that now since there are two types of individuals with different risk levels, they would respond to the donor priority rule differently by setting different cut-off costs, denoted by x_H and x_L, respectively, for high- and low-risk individuals. As a result, the arrival rate of type i donor patients is $\lambda_d^i(x_i) = p_i F(x_i)\,\theta_i$, and the arrival rate of type i non-donor patients is $\lambda_n^i(x_i) = p_i\,(1 - F(x_i))\,\theta_i$, for $i = H, L$. In addition, the total arrival rate of organs from both high- and low-risk organ donors is $\mu(x_H, x_L) = \sum_{i=L,H} p_i F(x_i)\phi n$.

Next, we represent the expected utility of each donor and non-donor. The total arrival rate of registered donors is $\lambda_d(x_H, x_L) = \sum_{i=L,H} \lambda_d^i(x_i) = \sum_{i=L,H} p_i F(x_i^*)\theta_i$. The donors are endowed with the priority to receive all the available organs, and thus the total supply rate of organs available to donors is $\mu_d(x_H, x_L) = \mu(x_H, x_L)$. Hence given x_H and x_L, a type i organ donor with a cost c has a net utility of

$$u_d^i(c, x_H, x_L) = \theta_i \beta T_p(x_H, x_L) - c,$$

where $T_p(x_H, x_L)$ is the aggregated post-transplantation life expectancy, that is, $T_p(x_H, x_L) = \frac{\sum_{i=L,H} p_i F(x_i) T_i}{\sum_{i=L,H} p_i F(x_i)}$.

The total arrival rate of non-donors is $\lambda_n(x_H, x_L) = \sum_{i=L,H} \lambda_n^i = \sum_{i=L,H} p_i (1 - F(x_i)) \theta_i$. The organs available to the non-donors are the surpluses after having satisfying the demands from the donors; thus the supply rate of organs available to non-donors is the $\mu_n(x_H, x_L) = \mu(x_H, x_L) - \lambda_d(x_H, x_L) = \sum_{i=L,H} p_i F(x_i) (\phi n - \theta_i)$. Hence given x_H and x_L, a type-i non-donor with a cost c has a net utility of

$$u_n^i(c, x_H, x_L) = \theta_i \left(\alpha \frac{\lambda_n(x_H, x_L) - \mu_n(x_H, x_L)}{\lambda_n(x_H, x_L)} d + \beta \frac{\mu_n(x_H, x_L)}{\lambda_n(x_H, x_L)} T_p(x_H, x_L) \right).$$

We characterize the equilibrium in the following proposition.

Proposition 3.3 *In equilibrium, the cut-off costs x_i^* satisfy:*

$$x_i^* = \theta_i \cdot (\beta T_p(x_H^*, x_L^*) - \alpha d) \cdot \frac{\sum_{j=L,H} p_j(\theta_j - F(c_j^*)\phi n)}{\sum_{j=L,H} p_j (1 - F(c_j^*))\theta_j} \textit{ for } i = H, L. \tag{3.9}$$

Proof. Consider a type-i individual with the cut-off cost $x_i^*, i = H, L$. In equilibrium, the individual is indifferent as to whether to become an organ donor or not, that is,

$$u_d^H(x_H^*, x_H^*, x_L^*) = u_n^H(x_H^*, x_H^*, x_L^*) \text{ and}$$
$$u_d^L(x_L^*, x_H^*, x_L^*) = u_n^L(x_L^*, x_H^*, x_L^*)$$

The above equation can be rewritten as $x_i^* = \theta_i \left(\beta T_p(x_H^*, x_L^*) - \alpha d \right) \frac{\lambda_n(x_H^*, x_L^*) - \mu_n(x_H^*, x_L^*)}{\lambda_n(x_H^*, x_L^*)}$, which can be further reorganized as (3.9). *Q.E.D.*

The following corollary immediately follows from Proposition 3.3:

Corollary 3.4 $\frac{x_H^*}{x_L^*} = \frac{\theta_H}{\theta_L} > 1.$

Corollary 3.4, in turn, gives the following corollary:

Corollary 3.5 $T_p(x_H^*, x_L^*) < \sum_{i=L,H} p_i T_i.$

Proof. By noticing that $T_p(x_H^*, x_L^*) = \frac{\sum_{i=L,H} p_i F(x_i^*) T_i}{\sum_{i=L,H} p_i F(c_i^*)}$ and $x_H^* < x_L^*$. *Q.E.D.*

Corollaries 3.4 and 3.5 reveal an unexpected consequence of introducing the donor priority rule: people with different health risks perceive disproportional level of attractiveness of becoming registered organ donors. More specifically, high-risk individuals are more likely to become organ donors. As a result, the average quality of the donated organs is lower than the average quality of organs from the overall population.

Next, we show that the introduction of the donor priority rule can lead to a lower social welfare.

Proposition 3.4 *The social welfare decreases after the introduction of the donor priority rule if*

$$\frac{\beta \cdot T_a - \alpha d}{\beta \cdot T_p(x_H^*, x_L^*) - \alpha d} > \frac{\sum_{i=L,H} p_i \theta_i}{F(0)\,\phi n}. \tag{3.10}$$

Proof. We use $\Delta U(c)$ to denote the utility change of a type-*i* individual with a cost of *c* due to the introduction of the donor priority rule.

$$\Delta U(c) = \begin{cases} \beta T_p(x_H^*, x_L^*) - \alpha d \frac{\sum_{i=L,H} p_i\theta_i - F(0)\phi n}{\sum_{i=L,H} p_i\theta_i} - \beta T_a \frac{F(0)\phi n}{\sum_{i=L,H} p_i\theta_i} & \text{if } c \le 0 \\ \beta T_p(x_H^*, x_L^*) - \alpha d \frac{\sum_{i=L,H} p_i\theta_i - F(0)\phi n}{\sum_{i=L,H} p_i\theta_i} - \beta T_a \frac{F(0)\phi n}{\sum_{i=L,H} p_i\theta_i} - c & \text{if } 0 < c \le x_i^* \\ \beta T_p(x_H^*, x_L^*) - \alpha d \frac{\sum_{i=L,H} p_i\theta_i - F(0)\phi n}{\sum_{i=L,H} p_i\theta_i} - \beta T_a \frac{F(0)\phi n}{\sum_{i=L,H} p_i\theta_i} - x_i^* & \text{otherwise,} \end{cases}$$

which implies that $\Delta U(c)$ is decreasing in c. This, in turn, gives one of the sufficient conditions for $W_D^1 - W_N^1 < 0$:

$$\beta T_p(x_H^*, x_L^*) - \alpha d \frac{\sum_{i=L,H} p_i\theta_i - F(0)\,\phi n}{\sum_{i=L,H} p_i\theta_i} - \beta T_a \frac{F(0)\,\phi n}{\sum_{i=L,H} p_i\theta_i} < 0,$$

which can be rewritten as (3.10). *Q.E.D.*

3.6 A Freeze Period Remedy

In the previous section, we have shown that the introduction of the donor priority rule can lead to a reduction in the social welfare due to the resultant imbalanced incentive structure to individuals with varying health statuses. In this section, we introduce a simple and easy-to-implement freeze period remedy. We show that the remedy can offset the quality distortion effect as a result of the donor priority rule. Therefore, when used in conjunction with the donor priority rule, this remedy can improve the social welfare by expanding the size of the donor registry without reducing the average quality of donated organs.

By introducing the freeze period mechanism, the donor priority rule is activated once a fixed period after the registration has elapsed. Assume that the fixed period is S, then as the time a person of type i getting sick and needing an organ satisfies the exponential distribution with mean $1/\theta_i$, the probability that he/she gets sick after the fixed period S is $e^{-\theta_i S}$. Therefore, assuming that people of type i choose to register as an organ donor if the cost is no more that $x_i^{\#}$, the arrival rate of the patients under the donor priority rule is

$$\lambda_p = \sum_{i=L,H} p_i F\left(x_i^{\#}\right)\theta_i e^{-\theta_i S},$$

and the arrival rate of the patients without donor priority is

$$\lambda_n = \sum_{i=L,H} p_i \theta_i \left(1 - F\left(x_i^{\#}\right) e^{-\theta_i S}\right)$$

The total arrival rate of organs is

$$\mu = \sum_{i=L,H} p_i F(x_i^{\#})\phi n.$$

The patients who have previously signed up as organ donors are endowed with the priority to receive all the available organs, i.e., total supply rate of organs available to donors is $\mu_p = \mu$, while the total supply rate of donors available to the patients without donor priority is

$$\mu_n = \mu - \lambda_p = \sum_{i=L,H} p_i F(x_i^{\#})\left(\phi n - \theta_i e^{-\theta_i S}\right).$$

The aggregated post-transplantation life expectancy for the patients receiving organs is

$$T_p(x_H^\#, x_L^\#) = \frac{\sum_{i=L,H} p_i F(x_i^\#) T_i}{\sum_{i=L,H} p_i F\left(x_i^\#\right)}.$$

Therefore, a type i organ donor with a cost c has a net utility of

$$u_d^i(c) = \theta_i \left(e^{-\theta_i S} \beta T_p(x_H^\#, x_L^\#) + \left(1 - e^{-\theta_i S}\right) \left(\alpha \frac{\lambda_n - \mu_n}{\lambda_n} d + \beta \frac{\mu_n}{\lambda_n} T_p(x_H^\#, x_L^\#) \right) \right) - c,$$

while a type-i non-donor with a cost c has a net utility of

$$u_n^i(c) = \theta_i \cdot \left(\alpha \frac{\lambda_n - \mu_n}{\lambda_n} d + \beta \frac{\mu_n}{\lambda_n} T_p(x_H^\#, x_L^\#) \right).$$

We characterize the equilibrium in the following proposition.

Proposition 3.5 *In equilibrium, the cut-off costs $x_i^\#, i = H, L$ satisfy:*

$$x_i^\# = \theta_i e^{-\theta_i S} \left(\beta T_p(x_H^\#, x_L^\#) - \alpha d \right) \frac{\sum_{i=L,H} p_i \left(\theta_i - F(x_i^\#) \phi n \right)}{\sum_{i=L,H} p_i \theta_i \left(1 - F\left(x_i^\#\right) e^{-\theta_i S} \right)} > 0$$

Proof. A type-i individual with the cut-off cost $c_i, i = H, L$, is indifferent as to whether to become an organ donor or not. In other words,

$$\begin{aligned} &\theta_i \left(e^{-\theta_i S} \beta T_p(x_H^\#, x_L^\#) + \left(1 - e^{-\theta_i S}\right) \left(\alpha \frac{\lambda_n - \mu_n}{\lambda_n} d + \beta \frac{\mu_n}{\lambda_n} T_p(x_H^\#, x_L^\#) \right) \right) - x_i^\# \\ &\quad = \theta_i \left(\alpha \frac{\lambda_n - \mu_n}{\lambda_n} d + \beta \frac{\mu_n}{\lambda_n} T_p(x_H^\#, x_L^\#) \right), \end{aligned}$$

which can be rearranged as

$$x_i^\# = \theta_i e^{-\theta_i S} \left(\beta T_p(x_H^\#, x_L^\#) - \alpha d \right) \frac{\lambda_n - \mu_n}{\lambda_n},$$

or

$$x_i^\# = \theta_i e^{-\theta_i S} \left(\beta T_p(x_H^\#, x_L^\#) - \alpha d \right) \frac{\sum_{i=L,H} p_i \left[\theta_i - F(x_i^\#) \phi n \right]}{\sum_{i=L,H} p_i \theta_i \left(1 - F\left(x_i^\#\right) e^{-\theta_i S} \right)}.$$

Q.E.D.

The following corollary immediately follows from Proposition 3.3:

Proposition 3.6 *Under the freeze period remedy with a freeze period of S,* $\frac{x_H^\#}{x_L^\#} = \frac{\theta_H e^{-\theta_H S}}{\theta_L e^{-\theta_L S}}$. *When* $S = \frac{\ln\left(\frac{\theta_H}{\theta_L}\right)}{\theta_H - \theta_L}$, *then* $\frac{x_H^\#}{x_L^\#} = \frac{\theta_H e^{-\theta_H S}}{\theta_L e^{-\theta_L S}} = 1$, *and* $T_p(x_H^\#, x_L^\#) = \sum_{i=L,H} p_i T_i$.

Proof. By noticing that $T_p(x_H^\#, x_L^\#) = \frac{\sum_{i=L,H} p_i F(x_i^\#) T_i}{\sum_{i=L,H} p_i F(x_i^\#)}$ and $x_H^\# = x_L^\#$. *Q.E.D.*

Under the freeze period specified in Proposition 3.6, the average donor quality of the registered organ donors is the same as that of the general population. In other words, the freeze period remedy eliminates the distorted incentives due to the donor priority rule. Furthermore, the donor priority rule ensures that the proportion of organ donors is higher than the case without the donor priority rule. These two aspects suggest that the social welfare improves under the joint effect of the donor priority rule and the freeze period remedy.

3.7 Concluding Remarks

This research is motivated by the widening gap between the increasing demand for donated organs and the steady size of the donor registry. We focus on analyzing the donor priority rule under which registered organ donors have priority over non-donors in receiving transplants. One would naturally expect that society is better off since more people would sign up for organ donation. Although this is indeed the case when individuals are homogeneous (§3.4), our analysis in §3.5 reveals a hidden incentive issue associated with the donor priority rule. Specifically, the policy will exert different incentives to the people with varying risk levels: less healthy individuals have a higher incentive than healthier ones to become registered organ donors. As a result, although the total organ donation rate is higher in response to the policy, the average quality of the donated organs can be lower due to the distorted incentives.

Our proposed remedy is to introduce a freeze period, which not only overcomes the loophole causing people to register as organ donors in the last minute before they need one, but also restores the overall quality of the donated organs. The reason is that the existence of a freeze period discourages both type of individuals from becoming potential organ donors, but it discourages high-risk individuals more than low-risk individuals. Thus, by appropriately choosing the length of the freeze period, the quality distorting effect introduced by the donor priority rule can be mitigated.

Opportunities for future research include extending the modeling of population heterogeneity as two group into multiple groups or as a continuous type. In either of these cases, we expect that our major insights remain unchanged, that is the donor priority rule can lead to a distorted pool of donated organs and hence reduce the social welfare. Nevertheless, we expect our proposed freeze period scheme to evolve into a scheme where the length of the freeze depends on individual characteristics (e.g., age). Further research could also explore the policy design problems specific to the characteristics of each type of organ.

Appendix A

Appendices for Chapter 1

A.1 Technical Proofs

Proof of Proposition 1.1. We first show that the physician's objective function $g(\mu, p) = p\lambda(\mu, p)$ is concave in p since $\partial^2 g(\mu, p)/\partial p^2 = -2p\beta^2\omega/[Q(\mu) - \pi - \beta(p - \pi)]^3 - 2\beta\omega/[Q(\mu) - \pi - \beta(p - \pi)]^2 < 0$. Solving the first-order condition gives the optimal service fee p^*, conditional on the service rate μ: $p^*(\mu) = \left\{\mu Q(\mu) - \mu(1-\beta)\pi - \sqrt{\mu\omega[Q(\mu) - (1-\beta)\pi]}\right\}/(\mu\beta)$, using which the physician's objective function can be rewritten as $g(\mu, p^*(\mu)) = \{r(\mu) + \omega - 2\sqrt{r(\mu)\omega}\}/\beta$, where $r(\mu) = \mu[Q_c + \alpha\ (\mu_c - \mu) - \pi(1-\beta)]$

Next, we show that $g(\mu, p^*(\mu))$ is unimodal in μ. Note that $r(\mu) \geq \mu[Q(\mu) - \pi(1-\beta)] > \mu[Q(\mu) - (\beta p + \pi(1-\beta))] \geq \mu\omega[\mu - \lambda(\mu, p^*(\mu))]^{-1} > \omega$. Hence the sign of

$$dg(\mu, p^*(\mu))/d\mu = \left(Q_c - \pi(1-\beta) + \alpha(-2\mu + \mu_c)\right) \cdot \left(\sqrt{r(\mu)\omega} - \omega\right) \cdot \left(\beta\sqrt{r(\mu)\omega}\right)^{-1}$$

is the same as that of $Q_c - \pi(1-\beta) + \alpha(-2\mu + \mu_c)$, which is positive when $\mu = 0$, decreases in μ, and turns negative when μ is large enough. $g(\mu, p^*(\mu))$ is therefore unimodal in μ. Equating the first-order derivative of $g(\mu, p^*(\mu))$ in terms of μ to zero gives $\mu^* = [Q_c + \alpha\mu_c - (1-\beta)\pi]/(2\alpha)$, which in turn yields $p^* = [Q_c + \alpha\mu_c - (1-\beta)\pi - 2\sqrt{\alpha\omega}]/(2\beta)$, and $\lambda^* = \lambda(\mu^*, p^*) = [Q_c + \alpha\mu_c - (1-\beta)\pi]/(2\alpha) - \sqrt{\omega/\alpha} = [Q_c + \alpha\mu_c - (1-\beta)\pi - 2\sqrt{\alpha\omega}]/(2\alpha)$. The expected waiting time can thus be determined given μ^* and $\lambda(\mu^*, p^*)$: $W^* = W(\mu^*, \lambda(\mu^*, p^*)) = [\mu^* - \lambda(\mu^*, p^*)]^{-1} = \sqrt{\alpha/\omega}$. ■

Proof of Proposition 1.2. We first recognize that $U(\mu, \lambda)$ is concave in μ as $\partial^2 U(\mu, \lambda)/\partial \mu^2 = -2\lambda\omega(\mu - \lambda)^{-3} < 0$ for any pair of (μ, λ) that satisfies $\mu > \lambda$. By the first-order condition (in terms of μ), we obtain the conditional expression of the optimal service rate: $\mu^S(\lambda) = \lambda + \sqrt{\omega/\alpha}$, using which the objective function can be rewritten as $-\alpha\lambda^2 + (\alpha\mu_c + Q_c - 2\sqrt{\alpha\omega})\lambda$, a concave function of λ. The first-order condition gives $\lambda^S = (Q_c + \alpha\mu_c)/(2\alpha) - \sqrt{\omega/\alpha}$, and hence $\mu^S = (Q_c + \alpha\mu_c)/(2\alpha)$. The expected waiting time is thus $W^S = W(\mu^S, \lambda^S) = (\mu^S - \lambda^S)^{-1} = \sqrt{\alpha/\omega}$. ■

Proof of Proposition 1.3. Similar to the proof of Proposition 1.1. ■

Proof of Corollary 1.5. We have two cases to consider depending on the size of $p_{\max}$ (cf. Proposition 1.3). Case (i): $p_{\max} > [Q_c + \alpha\mu_c - 2\sqrt{\omega\alpha} - (1 - \beta)\pi]/(2\beta)$. In this case, we have $\mu^* = [Q_c + \alpha\mu_c - (1 - \beta)\pi]/(2\alpha) < \mu^S = (Q_c + \alpha\mu_c)/(2\alpha)$. Case (ii): $p_{\max} \le [Q_c + \alpha\mu_c - 2\sqrt{\omega\alpha} - (1 - \beta)\pi]/(2\beta)$. Let $q_{\max} = \pi + \beta(p_{\max} - \pi)$. We have $q_{\max} \le [Q_c + \alpha\mu_c - 2\sqrt{\omega\alpha} + (1 - \beta)\pi]/2$. We then divide Case (ii) into two sub-cases depending on the size of $q_{\max}$: (a) If $(Q_c + \alpha\mu_c - 2\sqrt{\alpha\omega})/2 < q_{\max} \le [Q_c + \alpha\mu_c - 2\sqrt{\omega\alpha} + (1 - \beta)\pi]/2$, then $\mu^* = (Q_c + \alpha\mu_c - q_{\max} - \sqrt{\alpha\omega})/\alpha < (Q_c + \alpha\mu_c)/(2\alpha) = \mu^S$; (b) If $q_{\max} \le [Q_c + \alpha\mu_c - 2\sqrt{\alpha\omega}]/2$, then $\mu^* = (Q_c + \alpha\mu_c - q_{\max} - \sqrt{\alpha\omega})/\alpha \ge (Q_c + \alpha\mu_c)/(2\alpha) = \mu^S$. ■

Proofs of Propositions 1.4–1.5. Similar to the proof of Proposition 1.1. ■

Proof of Proposition 1.6. Suppose that $\mu_g^* \le \mu_b^*$, it suffices to prove that $\mu^* < \mu_g^*$. This consists of three steps.

First, we randomly pick a real number $\breve{\omega}$ that satisfies $\breve{\omega} > \max\{\underline{\omega}_b, \underline{\omega}_g\}$. We have

$$\frac{\rho_g \underline{\omega}_b + \rho_b \underline{\omega}_g}{\alpha\left(\rho_g + \rho_b\right)} < \frac{\breve{\omega}(\rho_g + \rho_b)}{\alpha\left(\rho_g + \rho_b\right)} = \frac{\breve{\omega}}{\alpha}. \tag{A.1}$$

Note that in a homogeneous system where all patients have a unit waiting cost of $\breve{\omega}$ and have type g insurance, the optimal service rate is μ_g^*, and each patient's total expected waiting time (including both the service time and the time in the queue) is $\sqrt{\alpha/\breve{\omega}}$. Since the total expected waiting time is longer than the expected service time, we have $\sqrt{\alpha/\breve{\omega}} > 1/\mu_g^*$, which can be reorganized as $\breve{\omega}/\alpha < (\mu_g^*)^2$. This, together with (A.1), gives

$$\frac{\rho_g \underline{\omega}_b + \rho_b \underline{\omega}_g}{\alpha\left(\rho_g + \rho_b\right)} < (\mu_g^*)^2. \tag{A.2}$$

Second, define $\bar{\mu} = \frac{\rho_g \mu_g^* + \rho_b \mu_b^*}{\rho_g + \rho_b}$. since $\mu_g^* \leq \mu_b^*$, we have $\bar{\mu} > \mu_g^*$,, which follows

$$\bar{\mu} + \sqrt{\bar{\mu}^2 - \frac{\rho_g \underline{\omega}_b + \rho_b \underline{\omega}_g}{\alpha \left(\rho_g + \rho_b\right)}} > \bar{\mu} > \mu_g^*. \tag{A.3}$$

Third, we rewrite the expression of μ^* as

$$\mu^* = \frac{\rho_g \underline{\omega}_b + \rho_b \underline{\omega}_g}{\alpha \left(\rho_g + \rho_b\right)} \cdot \left(\bar{\mu} + \sqrt{\bar{\mu}^2 - \frac{\rho_g \underline{\omega}_b + \rho_b \underline{\omega}_g}{\alpha \left(\rho_g + \rho_b\right)}}\right)^{-1},$$

which is lower than μ_g^* because on the right-hand side, the first part, by (A.2), is lower than $(\mu_g^*)^2$, and the second part, by (A.3), is lower than $1/\mu_g^*$. ■

Proof of Proposition 1.7. Depending on the relative magnitudes of $\underline{\omega}_b$, $\underline{\omega}_g$, and $\hat{\omega}$, two possible cases can arise: If $\underline{\omega}_g > \hat{\omega}$, that is, no type g patients' waiting costs exceed the threshold. Then, optimal service rate is $\mu^* = \lambda_b + \sqrt{(\underline{\omega}_b + \hat{\omega})/(2\alpha)}$, which is less than $(Q_c + \alpha\mu_c)/(2\alpha)$. If $\underline{\omega}_g \leq \hat{\omega}$, that is, some type g patients' waiting costs are below the threshold. Then, patient heterogeneity, again, makes it more desirable to admit fewer patients at any given time because of the increased average waiting cost as a result of increased access.

Case 1. $\underline{\omega}_g > \hat{\omega}$, that is, no type g patients' waiting costs exceed the threshold. In this case, only type b patients with waiting costs lower than $\hat{\omega}$ are admitted into the queue. The arrival rate is $\lambda_b = \Lambda_b \left(\hat{\omega} - \underline{\omega}_b\right)/\Delta\omega_b$. The social planner's problem is to choose the service rate μ and the cut-off time cost $\hat{\omega}$ to maximize the social welfare rate:

$$SW(\mu, \hat{\omega}) = \frac{\Lambda_b \left(\hat{\omega} - \underline{\omega}_b\right)}{\Delta\omega_b} \left[Q(\mu) - \frac{\underline{\omega}_b + \hat{\omega}}{2\left(\mu - \lambda_b\right)}\right]. \tag{A.4}$$

In this case, we can show in both that $SW(\mu, \hat{\omega})$ is concave in μ, and the optimal service rate $\mu^*(\hat{\omega}) = \lambda_b + \sqrt{(\underline{\omega}_b + \hat{\omega})/(2\alpha)}$. Substituting this intermediate result into (A.4), we can write the social welfare as a function of λ_b and $\hat{\omega}$:

$$SW(\lambda_b, \hat{\omega}) = \lambda_b \cdot \left[Q_c + \alpha\mu_c - \alpha\lambda_b - \sqrt{2\alpha(\underline{\omega}_b + \hat{\omega})}\right]. \tag{A.5}$$

Note that (A.5) is in essence a function of a single variable λ_b, since $\hat{\omega} = \Delta\omega_b/\Lambda_b \cdot \lambda_b + \underline{\omega}_b$. Solving the first-order condition gives

$$Q_c + \alpha\mu_c - \sqrt{2\alpha(\underline{\omega}_b + \hat{\omega})} - 2\alpha\lambda_b - \frac{\Delta\omega_b}{\Lambda_b} \cdot \sqrt{\frac{\alpha}{2(\underline{\omega}_b + \hat{\omega})}} = 0, \tag{A.6}$$

Since

$$0 < \frac{\Delta\omega_b}{\Lambda_b} \cdot \sqrt{\frac{\alpha}{2(\underline{\omega}_b + \hat{\omega})}} < \frac{\Delta\omega_b}{2\Lambda_b} \cdot \sqrt{\frac{\alpha}{\underline{\omega}_b}},$$

it follows from (A.6) that

$$\frac{Q_c + \alpha\mu_c}{2\alpha} - \frac{\Delta\omega_b}{4\Lambda_b\sqrt{\alpha\underline{\omega}_b}} - \sqrt{\frac{\underline{\omega}_b + \hat{\omega}}{2\alpha}} \le \lambda_b^* \le \frac{Q_c + \alpha\mu_c}{2\alpha} - \sqrt{\frac{\underline{\omega}_b + \hat{\omega}}{2\alpha}},$$

and hence

$$\frac{Q_c + \alpha\mu_c}{2\alpha} - \frac{\Delta\omega_b}{4\Lambda_b\sqrt{\alpha\underline{\omega}_b}} \le \mu^* = \lambda_b^* + \sqrt{\frac{\underline{\omega}_b + \hat{\omega}}{2\alpha}} \le \frac{Q_c + \alpha\mu_c}{2\alpha}.$$

This shows that the existence of patient heterogeneity essentially reduces the socially efficient service rate. The underlying explanation is that, as the arrival rate increases, the average waiting cost also increases. The social planner, therefore, admits fewer patients at any given time, and provides slower service for each patient accordingly.

Case 2. $\underline{\omega}_g \le \hat{\omega}$, that is, some type g patients' waiting costs are lower than the threshold. In this case, both types of patients with waiting costs lower than $\hat{\omega}$ are admitted into the queue. The choice of the admission control parameter $\hat{\omega}$ leads to arrival rates of $\lambda_i^S = (\hat{\omega} - \underline{\omega}_i)/\Delta\omega_i \cdot \Lambda_i, i = g, b$. The social planner chooses the service rate μ and the admission control parameter $\hat{\omega}$ to maximize the social welfare rate:

$$SW(\mu, \hat{\omega}) = \sum_{i=g,b} \lambda_i \cdot \left[Q(\mu) - \frac{\underline{\omega}_i + \hat{\omega}}{2} \cdot \frac{1}{\mu - \lambda_g - \lambda_b} \right],$$

where $\lambda_i = \Lambda_i \left(\hat{\omega} - \underline{\omega}_i\right)/\Delta\omega_i$ for $i = g, b$. The above equation suggests that patient heterogeneity, again, makes it desirable to admit fewer patients at any given time because of the increased average waiting cost as a result of increased access. ■

A.2 Analysis of Physician Type Uncertainty

To facilitate future analysis, define the function

$$V(\alpha) := Q_c + \alpha\mu_c - \pi\left(1 - \beta\right) - 2\sqrt{\alpha\omega}.$$

Notice that V is an increasing convex function if $\mu_c \geq \sqrt{\omega/\alpha}$. Next, we define the optimal service and revenue rates of a physician when he chooses price p and the patients believe him to be of type α:

$$\begin{aligned} \hat{\mu}(p,\alpha) &= \frac{V(\alpha)}{\alpha} - \frac{\beta}{\alpha}p + \sqrt{\frac{\omega}{\alpha}} \quad \text{for } \alpha \in [\alpha_L, \alpha_H] \\ \hat{g}(p,\alpha) &= \frac{V(\alpha)}{\alpha}p - \frac{\beta}{\alpha}p^2 \quad \text{for } \alpha \in [\alpha_L, \alpha_H]. \end{aligned}$$

Then, $\hat{g}(p,\alpha)$ is a concave function of p, which attains its maximum at $p^*(\alpha) = V(\alpha)/2\beta$. The optimal revenue rate at $p^*(\alpha)$ is equal to $\bar{g}(\alpha) := \hat{g}(p^*(\alpha), \alpha) = V^2(\alpha)/4\alpha\beta$, which is concave due to the Convex Maximum Theorem, cf. Carter (2001).

Proof of Corollary 9. Under the market equilibrium characterized in Proposition 1, the type α physician's optimal expected revenue rate can be written as $g(\mu^*, p^*, \alpha) = V^2(\alpha)/4\alpha\beta$, which is increasing in the skill level α since

$$\partial g(\mu^*, p^*; \alpha)/\partial\alpha = -\left(Q_c - \pi - \mu_c\alpha + \beta\pi\right)\left(Q_c - \pi + \mu_c\alpha + \beta\pi - 2\sqrt{\alpha\omega}\right)\left(1/4\alpha^2\beta\right)$$

is positive under Assumption 2. The optimal price $p^*(\alpha)$ increases because

$$\frac{dp^*(\alpha)}{d\alpha} = \frac{V'(\alpha)}{2\beta} = \nu - \frac{\omega}{\sqrt{\alpha\omega}} > 0,$$

where the last inequality follows from Assumption 1. Similarly, the optimal service rate decreases in α, implying an increase in the number of tests ordered with skill level. ■

Theorem A.1 *(Cho and Kreps 1987) In any monotonic, Spencian signaling game (cf. Spence 1973) with two sender types the Intuitive Criterion selects an equilibrium with the Riley outcome (cf. Riley 1979).*

Lemma A.1 *The signalling game described in Section 4.4 falls into the class of monotonic signaling games.*

Proof of Lemma A.1. Consider the following signalling game between two players: Player 1 (the physician) whose type α is privately-known and whose observable action is service price p) and Player 2 (competitive healthcare market representing the patients). Physician's type α is privately-known and her observable action is service fee p. The competitive healthcare market leaves patients indifferent between joining the queue and not (i.e., market clearing), therefore offering the physician a total payment $y = p\lambda\left(p\right) = V\left(\alpha\right) p/\alpha - \beta p^2/\alpha$. The payoffs are (u_1, u_2), where

$$u_1\left(\alpha, p, y\right) = y + \hat{g}(p, \alpha) \text{ and } u_2\left(\alpha, p, y\right) = -\left(y - \left(V\left(\alpha\right) p/\alpha - \beta p^2/\alpha\right)\right)^2.$$

Notice that this game possesses the following properties:

- The physician wants a higher payment y (monotonicity): For all $y' > y$, $u_1(\alpha, m, y') > u_1(\alpha, m, y)$.
- The patients' unique best response increases in their belief about physician's skill level α: For all p, $\partial\hat{g}(p, \alpha)/\partial\alpha > 0$
- It is relatively less costly for the high type physician to charge higher prices:

$$\frac{\partial u_1(\alpha, p, y)/\partial p}{\partial u_1(\alpha, m, p)/\partial p}$$

  is increasing in α.

Hence, we conclude that this is game is a monotonic signaling game with single-crossing (Spence 1973). ■

Proof of Proposition 1.8. Theorem A.1 and Lemma A.1 imply that in the signalling game, only the least-distortive separating (or Riley) equilibrium outcome satisfies the intuitive criterion. Hence, it suffices to consider the following maximization problem of the high type physician: Let $\alpha_H = \alpha + \delta$ and $\alpha_L = \alpha$. Then, the problem of type $\alpha + \delta$ physician is to choose a price p to signal its skill level by fully separating from type L:

$$\max_p \; 2\hat{g}(p, \alpha + \delta) \text{ subject to } \hat{g}(p, \alpha + \delta) + \hat{g}(p, \alpha) \leq 2\bar{g}\left(\alpha\right).$$

The Lagrangian for this problem is

$$L\left(p, \Lambda\right) = 2\hat{g}(p, \alpha + \delta) - \Lambda\left(\hat{g}(p, \alpha + \delta) + \hat{g}(p, \alpha) - 2\bar{g}\left(\alpha\right)\right).$$

The Kuhn-Tucker-Karush conditions yield

$$2\partial\hat{g}(p, \alpha + \delta)/\partial p = \Lambda(\frac{\partial\hat{g}(p, \alpha + \delta)}{\partial p} + \frac{\partial\hat{g}(p, \alpha)}{\partial p}), \tag{A.7}$$

and

$$\Lambda(\hat{g}(p, \alpha + \delta) + \hat{g}(p, \alpha) - 2\bar{g}(\alpha)) = 0. \tag{A.8}$$

If the IC constraint of the low type does not bind, we have $\Lambda = 0$, and part (i) of Proposition 8 follows from (A.7). If the IC constraint binds and $\Lambda > 0$, (A.7)–(A.8) yield:

$$\begin{aligned} \Lambda &= \frac{2\partial\hat{g}(p, \alpha + \delta)/\partial p}{\partial\hat{g}(p, \alpha + \delta)/\partial p + \partial\hat{g}(p, \alpha)/\partial p}, \text{ and} \\ -\left(\frac{\beta}{\alpha + \delta} + \frac{\beta}{\alpha}\right)\rho^2 &+ \left(\frac{V\left(\alpha + \delta\right)}{\alpha + \delta} + \frac{V\left(\alpha\right)}{\alpha}\right)\rho - \frac{V^2\left(\alpha\right)}{2\alpha\beta} = 0, \end{aligned}$$

the root of which is

$$\bar{p} = \frac{\left(\alpha V\left(\alpha + \delta\right) + \left(\alpha + \delta\right)V\left(\alpha\right)\right)}{2\beta\left(2\alpha + \delta\right)} + \frac{\sqrt{\left(\alpha V\left(\alpha + \delta\right) + \left(\alpha + \delta\right)V\left(\alpha\right)\right)^2 - \left(\left(\alpha + \delta\right)\left(2\alpha + \delta\right)V\left(\alpha\right)\right)^2}}{2\beta\left(2\alpha + \delta\right)}.$$

Substituting for $\alpha_h = \alpha + \delta$ and $\alpha_l = \alpha$ proves part (ii) of the proposition.

Finally, we derive the threshold skill level difference. First define the low type's deviation profit function $D\left(p,\alpha,\delta\right) = \hat{g}(p,\alpha+\delta) + \hat{g}(p,\alpha) - 2\bar{g}\left(\alpha\right)$, which is a quadratic function of high type's service fee p. Then, $D\left(p_h^*,\alpha,\delta\right) \leq 0$ if and only if its bigger root $\bar{p}$ is less than p_h^*. That is,

$$\frac{\left(\alpha V\left(\alpha+\delta\right)+\left(\alpha+\delta\right)V\left(\alpha\right)\right)}{2\beta\left(2\alpha+\delta\right)}+\frac{\sqrt{\left(\alpha V\left(\alpha+\delta\right)+\left(\alpha+\delta\right)V\left(\alpha\right)\right)^2-\left(\left(\alpha+\delta\right)\left(2\alpha+\delta\right)V\left(\alpha\right)\right)^2}}{2\beta\left(2\alpha+\delta\right)}=\frac{V\left(\alpha+\delta\right)}{2\beta}.$$

Hence, we define $\underline{\Delta\alpha}= \delta$ where $\bar{p} = p_h^*$. Arranging terms, the above inequality can be rewritten as $\alpha V^2\left(\alpha+\delta\right) \leq \left(\alpha+\delta\right)\left(V\left(\alpha+\delta\right)-V\left(\alpha\right)\right)^2$, which is non-negative only if$\sqrt{\alpha}V\left(\alpha+\delta\right) \leq \sqrt{\left(\alpha+\delta\right)}\left(V\left(\alpha+\delta\right)-V\left(\alpha\right)\right)$. Substituting for $V\left(\cdot\right)$, and arranging terms, we obtain the inequality

$$Q_c-\pi\left(1-\beta\right)+\mu_c\alpha-2\sqrt{\alpha\omega}\leq\mu_c\delta\left[\sqrt{\alpha_H/\alpha_L}-1\right]-2\sqrt{\alpha\omega}\left(\sqrt{\alpha_H/\alpha_L}-1\right)^2,$$

which defines $\underline{\Delta\alpha}$. ■

Appendix B

Appendices for Chapter 2

B.1 Technical Proofs

Proof of Lemma 2.1. Since $\partial\pi_S(Q^{FB}, Q_e)/\partial Q_e = p(1-\alpha)\beta\gamma\left[F((Q^{FB}-\gamma Q_e)/(1-\gamma)) - F(Q_e)\right] - (c_e - \beta c_r)$, we see that $\partial\pi_S(Q^{FB}, Q_e)/\partial Q_e\big|_{Q_e=Q^{FB}} = -(c_e-\beta c_r) < 0$, meaning that $Q_e^{FB} < Q^{FB}$.

□

Proof of Proposition 2.1(i). The retailer's marginal profit with respect to Q is

$$\partial\pi_R^{RS}(Q, Q_e)/\partial Q = (1-\psi)p \cdot \frac{\partial E[Z(Q, Q_e)]}{\partial Q} - (w + c_o) < (1-\psi)p - (w+c_o)$$

for any $Q > 0$ because

$$\partial E[Z(Q, Q_e)]/\partial Q = \alpha\bar{F}(Q) + (1-\alpha)\left[(1-\beta)\bar{F}\left(\frac{Q}{1-\gamma}\right) + \beta\bar{F}\left(\frac{Q-\gamma Q_e}{1-\gamma}\right)\right] < 1.$$

Hence we need to have $(1-\psi)p - (w+c_o) > 0$, or

$$\psi < 1 - (w + c_o)/p, \tag{B.1}$$

to rule out the case where $\partial\pi_R^{RS}(Q, Q_e)/\partial Q$ is negative for any $Q > 0$.

Next we analyze the manufacturer's marginal profit with respect to Q_e:

$$\partial\pi_M^{RS}(Q, Q_e)/\partial Q_e = \psi p \partial E[Z(Q, Q_e)]/\partial Q_e - (c_e - \beta c_r)$$

$$= \psi p(1-\alpha)\beta\gamma\left[F\left(\frac{Q-\gamma Q_e}{1-\gamma}\right) - F(Q_e)\right] - (c_e - \beta c_r)$$

$$< (1-\alpha)\beta\gamma[p-(w+c_o)] - (c_e - \beta c_r) \text{ (by (B.1))}$$

$$< 0 \text{ (by } (1-\alpha)\beta\gamma[p-(w+c_o)] < c_e - \beta c_r).$$

Therefore, the manufacturer will always choose $Q_e = 0$. □

Proof of Proposition 2.1(ii). Since $Q_e = 0$ for any ψ, it suffices to show that the retailer's order quantity Q decreases in ψ, which is true since $\partial \pi_R^{RS}(Q, Q_e)/\partial Q = (1-\psi)p\partial E[Z(Q,Q_e)]/\partial Q - (w + c_o)$ decreases in ψ. Since $Q < Q^{FB}$, we see that a higher ψ induces the retailer to choose a lower order quantity Q that deviates further from the first-best order quantity Q^{FB}. Therefore, the revenue sharing contract is strictly dominated by the wholesale price contract, a degenerate revenue sharing contract with $\psi = 0$. □

Proof of Lemma 2.2. Part (i) follows from (a) $\partial E[Z(Q,Q_e)]/\partial Q_e = (1-\alpha)\beta\gamma[F(\frac{Q-\gamma Q_e}{1-\gamma})$ $-F(Q_e)] \geq 0$, because $\frac{Q-\gamma Q_e}{1-\gamma} \geq \frac{Q_e-\gamma Q_e}{1-\gamma} = Q_e$, and (b) $\partial E[Z(Q,Q_e)]/\partial Q = \alpha\bar{F}(Q) + (1-\alpha)[(1-\beta)\bar{F}(\frac{Q}{1-\gamma}) + \beta\bar{F}(\frac{Q-\gamma Q_e}{1-\gamma})] > 0$. Part (ii) follows from $\partial^2 E[Z(Q,Q_e)]/\partial Q_e \partial Q = \frac{\beta\gamma(1-\alpha)}{1-\gamma} \cdot f\left(\frac{Q-\gamma Q_e}{1-\gamma}\right) > 0$. □

Proof of Proposition 2.2. In each of following two cases, we first derive and evaluate the manufacturer's marginal benefit from operating in the early production mode for an arbitrary combination of (Q, Q_e), and then apply the analysis to the case that $Q = Q^{FB}$ and $Q_e = Q_e^{FB}$.

(i) If $\kappa_1 Q_e + \kappa_2(Q - Q_e) < Q - Q_e$, then $\kappa_2 - \kappa_1 > 1 - (1-\kappa_2)Q/Q_e$. We have from (2.6), (2.7), and (2.9) that

$$\begin{aligned}
E[R_d(Q,Q_e)] = {} & \alpha\left[\int_0^{(1-\kappa_1)Q} \kappa_1 Q dF(\xi) + \int_{(1-\kappa_1)Q}^{Q} (Q-\xi) dF(\xi)\right] \\
& + (1-\alpha)\beta\Bigg[\int_0^{\frac{(1-\kappa_2)Q+(\kappa_2-\kappa_1-\gamma)Q_e}{1-\gamma}} [\kappa_1 Q_e + \kappa_2(Q-Q_e)]\, dF(\xi) \\
& + \int_{\frac{(1-\kappa_2)Q+(\kappa_2-\kappa_1-\gamma)Q_e}{1-\gamma}}^{\frac{Q-\gamma Q_e}{1-\gamma}} [Q - \gamma Q_e - (1-\gamma)\xi]\, dF(\xi)\Bigg] \\
& + (1-\alpha)(1-\beta)\left[\int_0^{\frac{(1-\kappa_2)Q}{1-\gamma}} \kappa_2 Q dF(\xi) + \int_{\frac{(1-\kappa_2)Q}{1-\gamma}}^{\frac{Q}{1-\gamma}} [Q - (1-\gamma)\xi]\, dF(\xi)\right].
\end{aligned}$$

The first-order derivative of $E[T_d]$ with respect to Q_e is

$$\partial E[T_d(Q,Q_e)]/\partial Q_e = w(1-\alpha)\beta\gamma\left[F\left(\frac{(1-\kappa_2)Q+(\kappa_2-\kappa_1-\gamma)Q_e}{1-\gamma}\right)-F\left(\frac{Q-\gamma Q_e}{1-\gamma}\right)\right]$$
$$-w(1-\alpha)\beta(\kappa_2-\kappa_1)F\left(\frac{(1-\kappa_2)Q+(\kappa_2-\kappa_1-\gamma)Q_e}{1-\gamma}\right). \qquad \text{(B.2)}$$

If $\kappa_1 \geq \kappa_2$, then the second term in the right-hand side of (B.2) is negative or zero. In this case, the marginal benefit from operating in the early mode to the manufacturer, captured by $(-\partial E[T_d(Q,Q_e)]/\partial Q_e)$, is lower than that to the supply chain, captured by $p \cdot \partial E[Z(Q,Q_e)]/\partial Q_e$. To see this, note that $\frac{(1-\kappa_2)Q+(\kappa_2-\kappa_1-\gamma)Q_e}{1-\gamma} > \frac{[1-(\kappa_2-\kappa_1)]Q_e+(\kappa_2-\kappa_1-\gamma)Q_e}{1-\gamma} = \frac{(1-\gamma)Q_e}{1-\gamma} = Q_e$, which gives

$$-\partial E[T_d(Q,Q_e)]/\partial Q_e \leq w(1-\alpha)\beta\gamma\left[F\left(\frac{Q-\gamma Q_e}{1-\gamma}\right)-F\left(\frac{(1-\kappa_2)Q+(\kappa_2-\kappa_1-\gamma)Q_e}{1-\gamma}\right)\right]$$
$$< p\cdot\partial E[Z(Q,Q_e)]/\partial Q_e = p(1-\alpha)\beta\gamma\left[F\left(\frac{Q-\gamma Q_e}{1-\gamma}\right)-F\left(Q_e\right)\right].$$

Therefore, any D-QF contract with $\kappa_1 \geq \kappa_2$ cannot coordinate the supply chain. However, when we choose $\kappa_2 > \kappa_1$, the second term in (B.2) is positive, meaning that the D-QF contract provides additional incentives for the manufacturer to operate in the early production mode. To coordinate the supply chain, (κ_1,κ_2) must satisfy $-\partial E[T_d(Q,Q_e)]/\partial Q_e = p\cdot\partial E[Z(Q,Q_e)]/\partial Q_e$ at $(Q,Q_e) = (Q^{FB},Q_e^{FB})$, which can be reorganized as $(\kappa_2-\kappa_1-\gamma)\cdot F\left(\frac{(1-\kappa_2)Q^{FB}+(\kappa_2-\kappa_1-\gamma)Q_e^{FB}}{1-\gamma}\right)+\gamma F(Q_e^{FB}) = \frac{(c_e-\beta c_r)(p-w)}{wp(1-\alpha)\beta}$.

(ii) If $\kappa_1 Q_e + \kappa_2(Q-Q_e) \geq Q-Q_e$, then $\kappa_2-\kappa_1 \leq 1-(1-\kappa_2)Q/Q_e$. We have from (2.6), (2.8), and (2.9) that

$$E[R_d(Q,Q_e)] = \alpha\left[\int_0^{(1-\kappa_1)Q}\kappa_1 Q dF(\xi)+\int_{(1-\kappa_1)Q}^{Q}(Q-\xi)dF(\xi)\right]$$
$$+(1-\alpha)\beta\left\{\int_0^{(1-\kappa_2)Q+(\kappa_2-\kappa_1)Q_e}\left[\kappa_1 Q_e+\kappa_2(Q-Q_e)\right]dF(\xi)\right.$$
$$\left.+\int_{(1-\kappa_2)Q+(\kappa_2-\kappa_1)Q_e}^{Q_e}(Q-\xi)dF(\xi)+\int_{Q_e}^{\frac{Q-\gamma Q_e}{1-\gamma}}\left[Q-\gamma Q_e-(1-\gamma)\xi\right]dF(\xi)\right\}$$
$$+(1-\alpha)(1-\beta)\left\{\int_0^{\frac{(1-\kappa_2)Q}{1-\gamma}}\kappa_2 Q dF(\xi)+\int_{\frac{(1-\kappa_2)Q}{1-\gamma}}^{\frac{Q}{1-\gamma}}\left[Q-(1-\gamma)\xi\right]dF(\xi)\right\}.$$

The first-order derivative of $E[T_d]$ with respect to Q_e is

$$\partial E[T_d]/\partial Q_e$$
$$= w(1-\alpha)\beta\gamma\left[F(Q_e) - F\left(\frac{Q-\gamma Q_e}{1-\gamma}\right)\right] - w(1-\alpha)\beta(\kappa_2-\kappa_1)F((1-\kappa_2)Q + (\kappa_2-\kappa_1)Q_e). \tag{B.3}$$

Similar to Case (i), we need $\kappa_2 > \kappa_1$ so that the contract can provide adequate incentive for the manufacturer to operate in the early production mode. To coordinate the supply chain, it is necessary for (κ_1, κ_2) to satisfy $-\partial E[T_d(Q, Q_e)]/\partial Q_e = p \cdot \partial E[Z(Q, Q_e)]/\partial Q_e$ at $(Q, Q_e) = (Q^{FB}, Q_e^{FB})$, which can be reorganized as $(\kappa_2 - \kappa_1) \cdot F\left((1-\kappa_2)Q^{FB} + (\kappa_2-\kappa_1)Q_e^{FB}\right) = \frac{(c_e-\beta c_r)(p-w)}{wp(1-\alpha)\beta}$. □

Proof of Proposition 2.3. Under the BLR contract, the manufacturer's and the retailer's profit functions are, respectively, as follows:

$$\pi_M^{BLR}(Q_e, Q) = bE[Z(Q, Q_e)] + [w - c_r - b - \rho w(1-\alpha)]Q - [c_e - \beta c_r - \rho w\beta(1-\alpha)]Q_e, \tag{B.4}$$
$$\pi_R^{BLR}(Q, Q_e) = (p-b)E[Z(Q, Q_e)] - [w + c_o - b - \rho w(1-\alpha)]Q - \beta\pi(1-\alpha)Q_e. \tag{B.5}$$

The procedure of deriving the optimal contract parameters consists of two steps. First, suppose that the retailer orders exactly Q^{FB}, we choose the contract parameters to ensure that the manufacturer would respond by setting $Q_e = Q_e^{FB}$. To implement the first-best solution, we need to make sure that the manufacturer's risk-benefit tradeoff is equivalent to that under the first-best scenario, that is,

$$\frac{b}{c_e - \beta c_r - \rho w\beta(1-\alpha)} = \frac{p}{c_e - \beta c_r}, \tag{B.6}$$

where b and $c_e - \beta c_r - \rho w\beta(1-\alpha)$ are the coefficients of $E[Z(Q, Q_e)]$ and Q_e in (B.4), and p and $c_e - \beta c_r$ are the coefficients of $E[Z(Q, Q_e)]$ and Q_e in (2.4).

Second, we analyze the retailer's decision in the way similar to the first step. We obtain the following equation by comparing (B.5) and (2.4):

$$\frac{p-b}{w+c_o-b-\rho w(1-\alpha)}=\frac{p}{c_r+c_o}. \tag{B.7}$$

Equations (B.6) and (B.7) jointly give the coordinating contract parameters. □

Remark: The above proof relies on the comparison of cost and revenue parameters, and is independent of γ. Therefore, Proposition 2.3 holds even when γ is a function of the quantity of on-time delivered units.

Proof of Proposition 2.4. Under the QF contract, the manufacturer's objective function is $\pi_M^{QF}(Q,Q_e)=(w-c_r)Q-(c_e-\beta c_r)Q_e-E[T_c(Q,Q_e)]$, where $E[T_c(Q,Q_e)]$ is the expected transfer payment from the manufacturer to the retailer. Hence the marginal benefit of early production to the manufacturer is $-\partial E[T_c(Q,Q_e)]/\partial Q_e$. In the first-best scenario, by comparison, the marginal benefit of early production is

$$p\partial E[Z(Q,Q_e)]/\partial Q_e=p(1-\alpha)\beta\gamma\cdot[F((Q-\gamma Q_e)/(1-\gamma))-F(Q_e)]. \tag{B.8}$$

Case (i). $0\le\kappa<1-Q_e/Q$: The expected returning quantity $E[R(Q,Q_e)]$, from Table 2.4, is

$$\begin{aligned}&\alpha\left[\int_0^{(1-\kappa)Q}\kappa Q dF(\xi)+\int_{(1-\kappa)Q}^{Q}(Q-\xi)dF(\xi)\right]\\&+(1-\alpha)\beta\left[\int_0^{\frac{(1-\kappa)Q-\gamma Q_e}{1-\gamma}}\kappa Q dF(\xi)+\int_{\frac{(1-\kappa)Q-\gamma Q_e}{1-\gamma}}^{\frac{Q-\gamma Q_e}{1-\gamma}}\left[Q-\gamma Q_e-(1-\gamma)\xi\right]dF(\xi)\right]\\&+(1-\alpha)(1-\beta)\left[\int_0^{\frac{(1-\kappa)Q}{1-\gamma}}\kappa Q dF(\xi)+\int_{\frac{(1-\kappa)Q}{1-\gamma}}^{\frac{Q}{1-\gamma}}\left[Q-(1-\gamma)\xi\right]dF(\xi)\right].\end{aligned}$$

Note that $0\le\kappa<1-Q_e/Q$ gives $Q_e<(1-\kappa)Q=\frac{(1-\kappa)Q-\gamma(1-\kappa)Q}{1-\gamma}<\frac{(1-\kappa)Q-\gamma Q_e}{1-\gamma}$. We hence have

$$\begin{aligned}-\partial E[T_c(Q,Q_e)]/\partial Q_e&=w(1-\alpha)\beta\gamma\left[F\left(\frac{Q-\gamma Q_e}{1-\gamma}\right)-F\left(\frac{(1-\kappa)Q-\gamma Q_e}{1-\gamma}\right)\right]\\&<w\cdot\partial E[Z(Q,Q_e)]/\partial Q_e\end{aligned} \tag{B.9}$$

It follows from (B.8) and (B.9) that the marginal benefit of early production to the manufacturer is strictly lower than that to the supply chain for any (Q, Q_e). Therefore, unless $Q_e^{FB} = 0$, the QF contract cannot provide the manufacturer with adequate incentive to operate in the early production mode even when the retailer chooses Q^{FB}.

Cases (ii). $1 - Q_e/Q \leq \kappa < 1$: Using a procedure similar to that in Case (i), we obtain

$$-\partial E[T_c(Q, Q_e)]/\partial Q_e = w(1-\alpha)\beta\gamma \cdot [F((Q-\gamma Q_e)/(1-\gamma)) - F(Q_e)]. \qquad \text{(B.10)}$$

By comparing (B.8) and (B.10), we see that $-\partial E[T_c(Q, Q_e)]/\partial Q_e = w/p \cdot p\partial E[Z(Q, Q_e)]/\partial Q_e < p\partial E[Z(Q, Q_e)]/\partial Q_e$. Hence, unless that $Q_e^{FB} = 0$, the manufacturer is not provided with adequate incentive to operate in the early production mode even when the retailer chooses Q^{FB}.

To summarize Cases (i) and (ii), the QF contract cannot coordinate the supply chain unless $Q_e^{FB} = 0$. □

Proof of Proposition 2.5. For the LR contract to coordinate the supply chain, we need two conditions. First, the LR contract needs to eliminate the double marginalization, that is, $w + c_o - \rho_{LR} w(1-\alpha) = c_r + c_o$. Second, the LR contract needs to offset the manufacturer's expected additional costs due to early production so that the manufacturer will be indifferent about the choice of Q_e, that is, $-(c_e - \beta c_r) + \rho_{LR} w(1-\alpha)\beta = 0$. These two equations jointly give the optimal contract parameters. □

Remark: Similar to Proposition 2.3, Proposition 2.5 holds even when γ is a function of the quantity of on-time delivered units.

B.2 Analysis of Traditional Contracts

B.2.1 Wholesale Price Contract

We first consider the wholesale price contract. The manufacturer's profit is $\pi_M^W(Q, Q_e) = wQ - c_r E[Q_r] - c_e Q_e = (w - c_r)Q - (c_e - \beta c_r)Q_e$, which gives $\partial \pi_M^W(Q, Q_e)/\partial Q_e = -(c_e - \beta c_r) < 0$. Hence the manufacturer has no incentive to operate in the early production mode and always chooses $Q_e = 0$. The problem is hence reduced to the scenario where the manufacturer has a single production mode.

B.2.2 Buyback Contract

Under a buyback contract, the manufacturer pays the retailer b for each unit of unsold product. The manufacturer's total buyback cost is therefore $b \cdot \{Q - E[Z(Q, Q_e)]\}$. The manufacturer's expected profit under this contract is $\pi_M^B(Q, Q_e) = (w - c_r)Q - (c_e - \beta c_r)Q_e - b \cdot \{Q - E[Z(Q, Q_e)]\}$. In the first-best scenario, given Q^{FB}, the manufacturer sets Q_e that maximizes $pE[Z(Q^{FB}, Q_e)] - (c_e - \beta c_r)Q_e$ (cf. (2.4)), which differs from $bE[Z(Q^{FB}, Q_e)] - (c_e - \beta c_r)Q_e$ under the buyback contract. Since $b < p$, the manufacturer's marginal return from early production is strictly lower than in the first-best scenario. Therefore, the manufacturer sets $Q_e < Q_e^{FB}$, meaning that the buyback contract is non-coordinating.

B.2.3 Quantity Discount Contract

Under a quantity discount contract, the retailer pays a lower wholesale price when the order quantity exceeds a certain threshold. We can show that, similar to the wholesale price contract, the quantity discount contract cannot motivate the manufacturer to operate in the early production mode, i.e., $Q_e = 0$.

B.2.4 Sales Rebate Contract

Under a sales rebate contract, the manufacturer provides the retailer with a rebate when the sales exceeds a certain threshold. Similar to the wholesale price contract, the sales rebate contract will always lead to $Q_e = 0$ because, while the sales rebate contract encourages the retailer to place a large order, it essentially gives the manufacturer a negative marginal benefit from engaging in early production.

B.3 Data Source

According to our communication with a production manager at a major vaccine manufacturer, the unit production costs for both early and regular production are roughly the same. We choose $c_r = c_e = \$3.0$ to be consistent with the influenza vaccine literature (e.g., Deo and Corbett 2009; Cho 2010). The wholesale price $w = \$12$ is chosen to be approximately the average unit price of influenza vaccine according to the CDC vaccine price list during the 2011–12 season (CDC 2011c). The retail price, estimated at \$18, is the approximate average quoted price we collected from several local pharmacies including Rite Aid, CVS, Walgreen, and Giant Eagle, \$28, less the average cost of administration, \$10, based on American Academy of Pediatrics (2007). We choose $\alpha = 0.80, \beta = 0.95$, and $\gamma = 0.50$ based on rough estimation, and thus complement our numerical study with a broad range of parameters $\alpha \in \{0.60, 0.70, 0.80\}, \beta \in \{0.90, 0.95, 0.99\}$, and $\gamma \in \{0.10, 0.30, 0.50\}$.

Bibliography

Abadie, A., S. Gay. 2006. The impact of presumed consent legislation on cadaveric organ donation: a cross-country study. *J. Health Econom.* **25** 599–620.

Adida, E., D. Dey, H. Mamani. 2011. Operational issues and network effects in vaccine markets. Working Paper.

Akan, M., O. Alagoz, B. Ata, F. Erenay, A. Said. 2012. A broader view of designing the liver allocation system. *Operations Res.* **60**(4) 757–770.

Alderman, L. 2011. The doctor will see you ... eventually. *New York Times* (August 2) D6.

American Academy of Pediatrics (AAP). 2007. Immunization financing: where is the breaking point? `http://www.aap.org/immunization/pediatricians/pdf/TaskForceWhitePaper.pdf`.

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

Arifoğlu, K., S. Deo, S. Iravani. 2012. Consumption externality and yield uncertainty in the influenza vaccine supply chain: Interventions in demand and supply sides. *Management Sci.* **58**(6) 1072–1091.

Arya, A., B. Mittendorf. 2004. Using return policies to elicit retailer information. *RAND J. Econom.* **35**(3) 617–630.

Ata, B., S. Tayur, A. Skaro. 2012. OrganJet: overcoming geographical disparities in access to deceased donor kidneys in the United States. Working paper.

Augier, M., J. G. March. 2011. *The Roots, Rituals and Rhetorics of Change: North American Business Schools After the Second World War*. Stanford University Press, Stanford, CA.

Baicker, K., E. S. Fisher, A. Chandra. 2007. Malpractice liability costs and the practice of medicine in the medicare program. *Health Affair.* **26**(3) 841–852.

Berwick, D. M., A. D. Hackbarth. 2012. Eliminating waste in US health care. *JAMA: J. Amer, Med. Asso.* **307**(14) 1513–1516.

Bruzzone, P. 2008. Religious aspects of organ transplantation. *Transplantation Proceedings* **40**(4) 1064–1067.

Cachon, G. 2003. Supply chain coordination with contracts. A. G. De Kok, S. C. Graves, eds. *Handbooks in Operations Research and Management Science, Vol. 11. Supply Chain Management: Design, Coordination, and Operation*. Elsevier Science, Amsterdam, The Netherlands.

Cachon, G., M. Lariviere. 2001. Contracting to assure supply: how to share demand forecasts in a supply chain. *Management Sci.* **47**(5) 629–646.

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

Cachon, G. P., F. Zhang. 2006. Procuring fast delivery: sole-sourcing with information asymmetry. *Management Sci.* **52**(6) 881–896.

Carrier, E. R., J. D. Reschovsky, M. M. Mello, R. C. Mayrell, D. Katz. 2010. Physicians' fears of malpractice lawsuits are not assuaged by tort reforms. *Health Affair.* **29**(9) 1585–1592.

Carter, M. 2001. *Foundations of Mathematical Economics*. MIT Press, Cambridge, Massachusetts.

Carter, M. W. 2002. Diagnosis: Mismanagement of resources. *OR/MS Today* **29**(2) 26–32.

Centers for Disease Control and Prevention (CDC). 2011a. Seasonal influenza vaccine supply and distribution in the United States. `http://www.cdc.gov/flu/about/qa/vaxdistribution.htm`.

Centers for Disease Control and Prevention (CDC). 2011b. Selecting the viruses in the seasonal influenza (flu) vaccine. `http://www.cdc.gov/flu/professionals/vaccination/virusqa.htm`.

Centers for Disease Control and Prevention (CDC). 2011c. CDC Vaccine Price List. `http://www.cdc.gov/vaccines/programs/vfc/cdc-vac-price-list.htm`.

Chick, S. E., S. Hasija, J. Nasiry. 2012. Incentive alignment and information elicitation from manufacturers for public goods procurement. Working paper.

Chick, S. E., H. Mamani, D. Simchi-Levi. 2008. Supply chain coordination and influenza vaccination. *Oper. Res.* **56**(6) 1493–1506.

Cho, I.-K., D. M. Kreps. 1987. Signaling games and stable equilibria. *Quart. J. Econom.* **102**(2) 179–221.

Cho, S.-H. 2010. The optimal composition of influenza vaccines subject to random production yields. *Manufacturing Service Oper. Management* **12**(2) 256–277.

Cho, S.-H., C. S. Tang. 2012. Advance selling in a supply chain under uncertain supply and demand. *Manufacturing Service Oper. Management*. Forthcoming.

Christensen, C. M., J. H. Grossman, J. Hwang. 2009. *The Innovator's Prescription: A Disruptive Solution for Health Care*. New York: McGraw-Hill.

Clement, D. 2012. Interview with Janet Currie. *The Region*, Federal Reserve Bank of Minneapolis (September) 8–19.

Coffey, R. M. 1983. The effect of time price on the demand for medical-care services. *J. Human Res.* **18**(3) 407–424.

Congressional Budget Office. 2008. Increasing transparency in the pricing of health care services and pharmaceuticals. *Economic and Budget Issue Brief* (June 5).

Cutler, D. M., M. McClellan. 2001. Is technological change in medicine worth it? *Health Affairs* **20**(5) 11–29.

Dai, T., S. Tayur, J. Rheinbolt, R. J. Noecker. 2012. Patient experience improvement at UPMC Eye Center. Tepper School of Business Teaching Case, Carnegie Mellon University, Pittsburgh, PA.

Debo, L., U. Rajan, S. K. Veeraraghavan. 2012. Signaling by price in a congested environment. working paper.

Debo, L., B. Toktay, L. V. Wassenhove. 2008. Queuing for expert services. *Management Sci.* **54**(8) 1497–1512.

Debo, L., S. K. Veeraraghavan. 2012. Equilibrium in queues under unknown service times and service value. Working paper.

Delmonico, F. L., Robert M. A., Nancy S.-H. et al. 2002. Ethical incentives—not payment—for organ donation. *N. Eng. J. Medicine* **346**(25) 2002–5.

Deo, S., C. J. Corbett. 2009. Cournot competition under yield uncertainty: The case of the U.S. influenza vaccine market. *Manufacturing Service Oper. Management* **11** 563–576.

Deo, S., I. Kolesov, S. Waikar. 2012. Optimizing flu vaccine planning at NorthShore University HealthSystem. Harvard Business School Case, Boston, MA.

Donate Life America. 2011. *National Donor Designation Report Card.* `http://donatelife.net`.

Donohue, K. L., 2000. Efficient supply contracts for fashion goods with forecast updating and two production modes. *Management Sci.* **46**(11) 1397–1411.

Dulleck, U., R. Kerschbamer. 2006. On doctors, mechanics, and computer specialists: the economics of credence goods. *J. Economic Literature* **44**(1) 5–42.

Durango-Cohen, E. J., C. A. Yano. 2006. Supplier commitment and production decisions under a forecast-commitment contract. *Management Sci.* **52**(1) 54–67.

Earn, D., J. Dushoff, S. A. Levin. 2002. Ecology and evolution of the flu. *Trends in Ecology & Evolution* **17**(7) 334–340.

Economist. 2010. Clear diagnosis, uncertain remedy. *The Economist* (Feb 18).

Esmail, N., M. Walker. 2003. *Waiting Your Turn: Hospital Waiting Lists in Canada* (13th Edition). Vancouver: The Fraser Institute.

Evans, R. G. 1974. Supplier-induced demand: some empirical evidence and implications. M. Perlman ed. *The Economics of Health and Medical Care*. London, U.K.: Macmillan, 162–201.

Federgruen, A., N. Yang. 2008. Selecting a portfolio of suppliers under demand and supply risks. *Oper. Res.* **56**(4) 916–936.

Feldstein, M. S. 1973. The welfare loss of excess health insurance. *J. Political Econom.* **81**(March–April) 251–280.

Frick, K. D., M. E. Chernew. 2009. Beneficial moral hazard and the theory of the second best. *Inquiry* **46**(2) 229–240.

Fukuda, K., D. O'Mara, J. A. Singleton. 2002. How the delayed distribution of influenza vaccine created shortages in 2000 and 2001. *Pharm. Ther.* **27** 235–242.

Gallup Organization. 2005. *National survey of organ and tissue donation attitudes and behaviors.* Princeton, NJ: Gallup Organization.

Goldhill, D. 2009. How American health care killed my father. *The Atlantic* (September).

Grady, D. 2011. Children on Medicaid shown to wait longer for care. *New York Times* (June 15, 2011).

Gravelle, H., L. Siciliani. 2008. Optimal quality, waits and charges in health insurance. *J. Health Econom.* **27**(3) 663–674.

Green, L. V. 2006. Queueing analysis in healthcare. Hall, R.W., ed. *Patient Flow: Reducing Delay in Healthcare Delivery*. New York: Springer.

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

Grout, J. R., D. P. Christy. 1993. An inventory model of incentives for on-time delivery in Just-In-Time purchasing contracts. *Naval Res. Logistics* **40** 863–877.

Gupta, S. 2012. More treatment, more mistakes. *New York Times* (July 31) A23.

Ha, A. Y., S. Tong. 2008. Revenue sharing contracts in a supply chain with uncontractible actions. *Naval Res. Logist.* **55**(5) 419–431.

Harvard Team. 1999. *Harvard Report: Improving Hong Kong's Health Care System: Why and for Whom?* Hong Kong SAR Food and Health Bureau. `http://www.fhb.gov.hk/en/press_and_publications/consultation/HCS.HTM`

Hassin, R., M. Haviv. 2003. *To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems*. Norwell, MA: Kluwer Academic Publishers.

Health Industry Distributors Association. 2011. *2010-2011 Influenza Vaccine Production & Distribution*.

Howard, D. H. 2002. Why do transplant surgeons turn down organs?: A model of the accept/reject decision. *J. Health Econom.* **21**(6) 957–969.

Hwang, W., Bakshi, N., DeMiguel, V. 2012. Efficiency of simple penalty contracts when suppliers can improve reliability. Working paper. London Business School.

IOM (Institute of Medicine). 2010. *The Healthcare Imperative: Lowering Costs and Improving Outcomes: Workshop Series Summary*. Washington, DC: The National Academies Press.

Jackson, C. 2009. Integrated design-to-delivery for fast fashion: on time, on trend, and on profit. Aberdeen Group, Boston, MA.

Johnson, E. J., D. Goldstein. 2003. Do defaults save lives? *Science* **302** 1338–1339.

Jung, K.-T. 1998. Influence of a per-visit copayment on health care use and expenditures: The Korean experience. *J. Risk Ins.* **65**(1) 33–56.

Kessler, J. B., A. E. Roth. 2012. Organ allocation policy and the decision to donate. *American Economic Review* **102**(5) 2018–2047.

Kissick, W. 1994. *Medicine's Dilemmas: Infinite Needs versus Finite Resources.* New Haven, CT: Yale University Press.

Kiviat, B. 2010. Can price shopping improve health care? *Time* (April 19).

Kleinrock, L. 1975. *Queueing Systems. Vol. I: Theory.* New York, NY: John Wiley & Sons.

Kostami, V., S. Rajagopalan. 2009. Speed quality tradeoffs in a dynamic model. University of Southern California working paper.

Lariviere, M. 1998. Supply chain contracting and coordination with stochastic demand. *Quantitative Models for Supply Chain Management*. eds, S. Tayur, R. Ganeshan and M. Magazine. Boston, MA: Kluwer.

Lavee, J., T. Ashkenazi, G. Gurman, D. Steinberg. 2009. A new law for allocation of donor organs in Israel. *The Lancet* **375**(9720) 1131–1133.

Lofgren, E., N. H. Fefferman, Y. N. Naumov, J. Gorski, E. N. Naumova. 2007. Influenza seasonality: underlying causes and modeling theories. *J. Virology* **81**(11) 5429–5436.

Mamani, H., E. Adida, D. Dey. 2012. Vaccine market coordination using subsidy. *IIE Transactions on Healthcare Systems Engineering* **2**(1) 78–96.

Matthews, J. T. 2006. Egg-based production of influenza vaccine: 30 years of commercial experience. *The Bridge* **36**(3) 17–24.

Miller, C. C. 2010. Bringing comparison shopping to the doctor's office. *New York Times* (June 10) B1.

Mold, J. W., R. M. Hamm, L. H. McCarthy. 2010. The law of diminishing returns in clinical medicine: how much risk reduction is enough? *J. Amer. Board Fam. Med.* **23** 371–375.

Newhouse, J. P. 1978. Insurance benefits, out-of-pocket payments, and the demand for medical care: a review of the literature. RAND: Santa Monica, CA.

Newhouse, J. P. 1992. Medical care costs: how much welfare loss? *J. Economic Perspectives* **6**(3) 3–21.

Newhouse, J. P. 2002. *Pricing the Priceless: A Health Care Conundrum*. Cambridge: MIT Press.

Nowalk, M. P., R. K. Zimmerman, S. M. Cleary, R. D. Bruehlman. 2005. Missed opportunities to vaccinate older adults in primary care. *J. Amr. Board Fam. Pract.* **18**(1) 20–27.

O'Mara, D., K. Fukuda, J. A. Singleton. 2003. Influenza vaccine: ensuring timely and adequate supply. *Infect. Med.* **20** 548–554.

Organ Donation Taskforce. 2008. The potential impact of an opt out system for organ donation in the UK: an independent report from the Organ Donation Taskforce. Department of Health, U.K.

Organdonor.gov. 2012. The need is real: data. Accessed on April 15, 2012. `http://organdonor.gov/about/data.html`

Paç, M., S. Veeraraghavan. 2012. Strategic diagnosis and pricing in expert services. Wharton Working Paper.

Padmanabhan, V., I. P. L. Png. 1997. Manufacturer's returns policies and retail competition. *Marketing Sci.* **16**(1) 81–94.

Pasternack, B. A. 1985. Optimal pricing and return policies for perishable commodities. *Marketing Sci.* **4**(2) 166–176.

Pauly, M. V. 1980. *Doctors and Their Workshops: Economic Models of Physician Behavior.* Chicago, IL: The University of Chicago Press.

Phelps, C. E., J. P. Newhouse. 1974. Coinsurance, the price of time, and the demand for medical services. *Rev. Econom. Stat.* **56**(3) 334–342.

Pines, J. M., Z. F. Meisel. 2011. Why doctors order too many tests (It's not just to avoid lawsuits). *Time* (February 25).

Pinker, E. J. 2012. Can OR/MS be a change agent in healthcare? *Manufacturing Service Oper. Management* **14**(4) 495–511.

Plambeck, E., T. Taylor. 2005. Sell the plant? The impact of contract manufacturing on innovation, capacity and profitability. *Management Sci.* **51**(1) 133–150.

Prottas, J. M. 1983. Encouraging altruism: Public attitudes and the marketing of organ donation. *Milbank Memorial Fund Quarterly. Health and Society* **61**(2) 278–306.

Rao, V. M., D. C. Levin. 2012. The overuse of diagnostic imaging and the Choosing Wisely initiative. *Annals Intern. Med.* **157**(8) 574–576.

Rheinbolt, J., R. J. Noecker. 2009. Personal communication.

Rheinbolt, J. 2012. Personal communication.

Riley, J. 1979. Informational equilibrium. *Econometrica* **47**(2) 331–359.

Singleton, J. A. 2011. Influenza vaccination distribution and coverage, United States, 2010-11 and 2011-12 seasons. Presentation at Meeting of the ACIP (October 26).

Sirovich, B. E., S. Woloshin, L. M. Schwartz. 2011. Too little? Too much? Primary care physicians' views on US health care: a brief report. *Arch. Intern. Med.* **171**(17) 1582–1585.

Sorensen, R., J. Grytten. 1999. Competition and supplier-induced demand in a health care system with fixed fees. *Health Econom.* **8**(6) 497–508.

Spence, A. M. 1973. Job market signaling. *Quart. J. Econom.* **87**(3) 355–374.

Studdert, D. M., M. M. Mello, A. A. Gawande, T. K. Gandhi, A. Kachalia, C. Yoon, A. L. Puopolo, T. A. Brennan. 2006. Claims, errors, and compensation payments in medical malpractice litigation. *N. Engl. J. Med.* **354**(19): 2024–2033.

Su, X., S. A. Zenios. 2004. Patient choice in kidney allocation: the role of the queueing discipline. *Manufacturing Service Oper. Management* **6**(4) 280–301.

Su, X., S. A. Zenios. 2006. Recipient choice can address the efficiency-equity trade-off in kidney transplantation: a mechanism design model. *Management Sci.* **52**(11) 1647–1660.

Taylor, T. A. 2002. Supply chain coordination under channel rebates with sales effort effects. *Management Sci.* **48**(8) 992–1007.

Taylor, T. A. 2001. Channel coordination under price protection, midlife returns, and end-of-life returns in dynamic markets. *Management Sci.* **47**(9) 1220–1234.

Taylor, T. A., E. L. Plambeck. 2007. Simple relational contracts to motivate capacity investment. *Manufacturing Service Oper. Management* **9**(1) 94–113.

Taylor T. A., W. Xiao. 2009. Incentives for retailer forecasting: rebates vs. returns. *Management Sci.* **55**(10), 1654–1669.

Teresi, D. 2012. What you lose when you sign that donor card. *Wall Street Journal* (March 10) C3.

Tong, C., S. Rajagopalan. 2012. Pricing and operational performance in discretionary services. Working paper.

Tsay, A. 1999. Quantity-flexibility contract and supplier-customer incentives. *Management Sci.* **45**(10) 1339–1358.

Vaccines and Related Biological Products Advisory Committee (Committee). 2007. Transcripts and other meeting documents. `http://www.fda.gov/cber/advisory/vrbp/vrbpmain.htm.`

Veeraraghavan, S. K., Debo, L. 2009. Joining longer queues: information externalities in queue choice. *Manufacturing Service Oper. Management* **11**(4) 543–562.

Walshe, K., T. G. Rundall. 2001. Evidence-based management: from theory to practice in health care. *Milbank Quarterly* **79**(3) 429–457.

Wang, X., L. G. Debo, A. Scheller-Wolf, S. F. Smith. 2010. Design and analysis of diagnostic service centers. *Management Sci.* **56**(11) 1873–1890.

Welch, H. G. 2012. If you feel O.K., maybe you are O.K. *New York Times* (February 28) A25.

White House. 2009. Remarks by the President at the annual conference of the American Medical Association. `http://www.whitehouse.gov`

Williams, D. 2005. The influenza vaccine supply chain: structure, risk, and coordination. Master's Thesis, Massachusetts Institute of Technology, Cambridge, MA.

Zelder, M. 2000. Spend more, wait less? The myth of underfunded Medicare in Canada. *Fraser Forum* (August Special Issue). Vancouver: The Fraser Institute.

Zenios, S. A. 1999. Modeling the transplant waiting list: A queueing model with reneging. *Queueing Systems* **31** 239–251.