Private Health Service, Public Waiting Time and Patient Welfare: Theoretical Modeling and Empirical Evidence

by

Qu Qian

B.Sc., Nanjing University, 2001

M.Sc., Georgia Institute of Technology, 2003

M.Sc., National University of Singapore, 2003

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES
(Business Administration)

THE UNIVERSITY OF BRITISH COLUMBIA
(Vancouver)

March 2012

Abstract

Motivated by health reform debates and policy changes in Canada and other OECD countries, we study how the private and public health care impact on the public health waiting time and more generally the welfare of patients. This thesis encompasses theoretical and empirical research.

In Chapter 2, we develop a theoretical model and then empirically test the association between allowing private care financing and public waiting time using joint replacement surgery data of nine Canadian provinces. Two policies that induce private care financing are tested. The empirical results suggest that both policies are associated with shorter public waiting times. This work contributes to the existing literature by providing an empirical analysis of the relationship between private care financing and public health waiting time under a unique institutional setting.

In Chapter 3, we investigate the effect of physician dual practice on public waiting time and patient welfare. Motivated by Manitoba's cataract surgery evidence, our study shows that the public waiting time difference existing between dual-practice physicians and public-only physicians can be explained by service quality differentiation. Patients with lower time costs would have a longer waiting time if physician dual practice were allowed. But some of these patients could be better off by receiving a service of higher quality induced by allowing physician dual practice. This work contributes to the limited literature of physician dual practice.

In Chapter 4, we study the use of tax or subsidy on private care to improve income redistribution given that the public health system is financed by a head tax. When the utilization of the public health system is low, the health planner should subsidize private care to induce patients with higher time costs to the private sector. The production cost of public care would then be reduced and so would be the head tax that everyone pays for. We show that the optimal tax/subsidy decision improves income redistribution when the utilization of public health system is either high or low. This work contributes to the literature of public provision of private goods on income redistribution.

Preface

This preface provides a statement of co-authorship for the essays contained in this thesis.

The research topics of Chapter 2 and Chapter 3 were identified by Profs. Hong Chen and Anming Zhang. I did most of the analysis, preparation and revision of the essays under close supervision and guidance of Profs. Hong Chen and Anming Zhang. Profs. Hong Chen and Anming Zhang also provided editorial support throughout the process of revision. The data collection for Chapter 2 was approved by UBC Behavioral Research Ethics Board on June 9, 2009. The UBC Behavioral Research Ethics Board Certificate of Approval number is H09-01361.

The research topic of Chapter 4 was identified by myself. The research project was undertaken with Profs. Hong Chen and Anming Zhang jointly. I did most of the analysis, preparation and revision of the essay. Profs. Hong Chen and Anming Zhang provided critical advices and suggestions to the revisions of the essay.

Table of Contents

List of Tables

List of Figures

Acknowledgment

This thesis would not have been what it is without the encouragement and generous support from many individuals. I am grateful to every person who has provided his/her support to my research and study and who has contributed in some way to the process leading to my thesis.

I would like to express my deepest gratitude to my supervisors, Professor Hong Chen and Professor Anming Zhang, for their patience, inspirations and guidance throughout my doctoral study. This thesis would not have been its current form without their continuous support and advices. Being great mentors, they have taught me how to nurture research interests and how to conduct productive research. They are examples of accomplished scholars for me to follow. I am fortunate to have Professor Steven Shechter and Professor Ilan Vertinsky to serve in my thesis committee. Their encouragement and constructive critiques have contributed greatly to the development of my thesis. Furthermore, I would also like to thank other OPLOG faculty members who never hesitate to share their insights with me in various occasions.

Let me also thank a few individuals who have supported my research and study in the past six years. I should thank Canadian Institute for Health Information and Ms. Shirley Shen for providing me the joint replacement data. I would like to thank Elaine Cho who has been helping me all along my doctoral study in Sauder School of Business. I would like to thank Michelle Medalla who was always there to support me, in particular to my job application. Furthermore, many thanks to UBC and NSERC for their continuous financial support to my doctoral study.

Last but not least, I want to thank my parents and wife. I am forever indebted to their unconditional understanding, support and encouragement. It is to them that I dedicate this work.

Dedication

To my parents, my wife and my lovely daughter.

Chapter 1

Introduction

We begin with the motivation of this thesis and then discuss how we design the thesis to address the research questions in Section 1.1. We then present a brief overview of the three essays included in this thesis in Section 1.2. We conclude this chapter with an outline of the thesis in Section 1.3.

1.1 Motivation and Research Design

This thesis is motivated by health reform debates in Canada and other OECD countries regarding the long waiting times in the public health system (Sanmartin et al., 2000). Debates in Canada are heightened by a series of waiting time related lawsuits.[1] Among these lawsuits, the ruling of Chaoulli vs. Quebec[2] turns out to be the most controversial ruling of Supreme Court of Canada to date. Physician Jacques Chaoulli, together with joint replacement patient George Zeliotis, sued Quebec government for not allowing patients to purchase private insurance to cover their private care expenses in face of long public waiting times. In 2005, the Supreme Court of Canada ruled that Quebec's ban on private health insurance on medically necessary services violates the Quebec Charter of Human Rights and Freedoms. After the ruling, Quebec government was forced to set up waiting time guarantee for joint replacement surgeries and allow patients to purchase private health insurance to pay for their private care should the waiting time guarantee be exceeded (Prémont, 2007).

[1] These lawsuits may include, but not limited to, Chaoulli vs. Quebec in 2005 (joint replacement), Murray vs. Alberta in 2006 (joint replacement), McCreith and Holmes vs. Ontario in 2007 (cancer).

[2] See Chaoulli v. Quebec (Attorney General), [2005] 1 S.C.R. 791, 2005 SCC 35.

Many people believe that this ruling could lead to fundamental structural changes in the way Canadian provinces deliver health services.

Similar to Canada, long waiting time of essential or elective services in the public health system is a critical issue in many OECD countries in the face of rising demand and costs, limited public health budget and the advancement of technologies (Canadian Institute for Health Information, 2008b). Because long public waiting time is the outcome of the interplay of many factors, the root causes of the problem are far from being clear. To tackle the waiting time problem, two competing solutions have been proposed in Canada's health reform debates (Sanmartin et al., 2000). Some people see long public waiting time as a strong call for more resources to re-strengthen the public health system (Quesnel-Vallee et al., 2006), so the solution is to increase the public health funding (Kondro, 2007). On the other hand, some other people believe that the public health system alone would not resolve the problem. These people argue that the waiting time problem roots in the failure of Canada's single payer system – patients do not have a choice in the face of long public waiting times. Supporters of this rationale propose to broaden the role and scope of private care (Esmail and Walker, 2008). Nevertheless, most of the current public versus private debates are characterized much more by claims and counterclaims from competing ideological bases than evidence (Tuohy et al., 2004). We are motivated to provide a better understanding on these controversial issues through theoretical modeling and empirical research.

The relation between public and private sectors could be structured very differently in different settings (Tuohy et al., 2004). Most OECD countries adopt a universal health care system, which normally consists of a public sector that provides the basic and necessary medical services and a parallel supplemental private sector. However, the role and form of the private sector vary greatly across countries. For instance, in many Canadian provinces, the private sector is only allowed to cover pharmaceutics and non-medically necessary services (Flood and Archibald, 2001); While in many European countries, the private sector is allowed to cover medically necessary services. To respect these institutional differences, the literature of health economics on private care generally focuses on two dimensions. The first dimension is related to the financing of private care (e.g., Besley et al. (1999); Hurley et al. (2001); Tuohy et al. (2004); Siciliani and Hurst (2005); Willcox et al. (2007)). Research topics of this dimension include, but not limited to, the source of private care financing, the relation between public care financing and private care financing, and the impact of private care financing on the public health system. Literature on this dimension mainly consists of empirical studies, such as case studies that compare the strategies of private care financing across different countries. The second dimension is related to the provision of private care (e.g., Iversen (1993, 1997); Olivella (2003); Duckett (2005); Derrett et al. (2009) and papers listed in Eggleston and Bir (2006)). Research topics of this dimension include, but not limited to, the delivery of private care, the existence of a private sector and its impact on public health system. Literature on this dimension mainly consists of theoretical research. As shown in Table 1.1, the three essays included in this thesis cover both dimensions.

Table 1.1: The coverage of topics on private care

	Empirical	*Theoretical*
Financing	Chapter 2	Chapter 2 / Chapter 4
Provision	×	Chapter 3

As this thesis is motivated by Canada's health reform debates, the research topics and design of this thesis also reflect the focus of the debates. First, Canada's health reform debates center at the financing of private care. This is because in Canada, the health providers are not salaried employees of public hospitals. Instead, they are self-employed professionals who are usually paid by social insurance on a fee-for-service basis (Flood and Archibald, 2001), i.e., public health services are delivered privately. Therefore, the central issue of the public versus private debates is whether or not to allow private financing of medically necessary services. In both Chapter 2 and Chapter 4, we focus on this issue by discussing the regulatory policies of private care financing. Second, Canada's health reform debates focus on medically necessary services. Some medically necessary services, particularly the five priority areas targeted by the governments for waiting time reduction, are considered as privatizable. Waiting time problems of these five areas are covered excessively by media, which builds up the tension between patients and public health authorities. The health services chosen for this thesis are of the five priority areas: Chapter 2 discusses the waiting time problem of joint replacement surgeries; Chapter 3 is motivated by the empirical evidence of cataract surgeries in Manitoba. Third, in addition to the financing of private care, another controversial issue in Canada's health reform debates is related to physician's status of practice. In Canada, physicians need to choose opt-in status in order to receive payments from social insurance. The opt-in status prevents physicians from seeing private patients and receiving private payments. The dismantlement of this status disincentive is a controversial proposal in the health reform plans of some provinces. This thesis also discusses the topic of physician's status of practice: Chapter 3 studies the phenomenon of physician dual practice in Manitoba. These research topics for this thesis allow us to focus on the core of the debates.

We resort to both empirical and theoretical methodologies to investigate the chosen research topics. Either type of research has its strengths and limitations. Theoretical research allows us to investigate a topic in a general setting without subject to the availability of data. Nevertheless, due to the complexity of health care system, certain assumptions have to be made to simplify the analysis. Empirical research has the advantage of providing evidence-based answers to hypotheses that are otherwise undetermined in theoretical research. By including both empirical and theoretical research in this thesis, we are able to provide a well-rounded understanding of the issues raised in the public versus private debates.

In the following section, we provide a brief overview of the three essays included in this thesis. For each essay, we would discuss the research questions to be addressed, the research methodologies used, the main results and the contributions to literature.

1.2 Overview of the Thesis

1.2.1 Chapter 2: Private Care Financing and Public Waiting Time

Chapter 2 is an empirical study that investigates the following research question: would allowing privately funded health care reduce the public waiting time? This empirical study focuses on the dimension of private care financing. Two policies that induce private care financing are tested using joint replacement surgery waiting time data. Joint replacement surgery is one of the target areas for waiting time reduction in Canada. Compared to other empirical studies that use cross country data, the cross province data used in this study are more homogeneous in terms of patient demographics and institutional characteristics. Therefore, we are in a better position to test the association between private care financing and public waiting time.

The main hypothesis for test is derived from a demand and supply equilibrium by modeling the provision of public health services. To test the policy effects, ideally one would expect the data set to have a panel structure so that the panel data technique, i.e., Fixed-effects model, can be used to control for the heterogeneity of cross section units. However, the policies are fixed in time series, so we have to resort to general linear model and random-effects model to draw conclusions. Robust tests are conducted to access the robustness of the results.

This study shows that policies that induce private care financing are associated with shorter public waiting times. This work provides empirical evidence to one of the key issues in the public versus private debates in Canada. This work contributes to the existing literature by providing an empirical research that is based on Canada's institutional setting.

1.2.2 Chapter 3: Physician Dual Practice, Public Waiting Time and Patient Welfare

Chapter 3 is a theoretical study that discusses the phenomenon of physician dual practice. This study focuses on the dimension of private care provision. This study is motivated by the observations of waiting time differences existing between dual-practice physicians and public-only physicians in Manitoba cataract surgeries. We aim to provide a possible explanation for these waiting time differences by considering two effects that are induced by allowing physician dual practice - service quality differentiation and patient prioritization. By considering these two effects jointly, we are able to investigate the impact of physician dual practice on patient's waiting time and welfare. We show that allowing physician dual practice would lengthen the waiting time of patients with lower time costs. However, these patients could be better off as they receive a service of higher quality.

Although physician dual practice is a common phenomenon in many OECD and developing countries, the literature on physician dual practice is limited and recent (Eggleston and Bir, 2006). The objective of this study is more to provide managerial insights to a specific application than to provide a general model. Therefore, the modeling assumptions are mainly implied in or supported by Manitoba's empirical evidence.

1.2.3 Chapter 4: Tax or Subsidy on Private Care and Income Redistribution

Chapter 4 is a theoretical study that discusses the use of tax or subsidy on private care to improve income redistribution. This study focuses on the dimension of private care financing. The research questions of this chapter are motivated by the findings of Chapter 2. Chapter 2 shows that providing subsidy to private care is associated with shorter public waiting times. This chapter looks at the same issue from a different but related perspective: Under what circumstances does providing subsidy to private care improve income redistribution? Existing literature has shown that public provision of private goods could improve income redistribution even if the public provision is financed by a head tax. We aim to extend the existing literature by taking tax or subsidy decision into account.

We extend the model of Hoel and Sáther (2003) to study the public health planner's tax or subsidy decision. The objective of income redistribution is modeled by assigning welfare weights to different patient types. The analytical results derived from $M/M/1$ queue are substantiated by the numerical results of $M/G/1$ queue. This study shows that the health planner should subsidize private care when the utilization of the public health system is low. Subsidy to private care induces patients with higher time costs to the private sector. Therefore, the production cost of public health service would be reduced and so would be the head tax that everyone pays for. Additionally, this study shows that the optimally designed tax or subsidy rate improves income redistribution when the utilization of public health system is either high or low.

1.3 Outline of the Thesis

The remainder of the thesis is organized as a series of chapters. At the beginning of every chapter, we motivate the research questions in discussion and examine the related literature. We then present our analysis and results. We conclude each chapter with a summary of the main findings. The Conclusion chapter summarizes the main results of this thesis and discusses some ongoing work closely related to this thesis and the possible extensions for future research. Following them is the bibliography for all chapters. The mathematical proofs for each chapter are placed in the appendices at the end of the thesis.

Chapter 2

Would Allowing Privately Funded Health Care Reduce the Public Waiting Time? Empirical Evidence from Canadian Joint Replacement Surgery Data

2.1 Introduction

Long waiting times in the public health system have been a source of public concern in many OECD countries (Siciliani and Hurst, 2005), and therefore a target for policy initiatives. In Canada, long waiting times for elective surgeries have led to several famous lawsuits[1] and nationwide health reform debates. In 2005, physician Jacques Chaoulli, together with joint replacement patient George Zeliotis, sued Quebec government for banning the purchase of private insurance to cover private care expenses when patients were not able to obtain timely access to public health care through Medicare, Canada's single-payer public health system. The Supreme Court's ruling turns out to be highly contentious (Quesnel-Vallee et al., 2006). The ruling forced Quebec government to change its policies towards public waiting time

[1] These lawsuits include Chaoulli vs. Quebec in 2005 (joint replacement), Murray vs. Alberta in 2006 (joint replacement), and McCreith and Holmes vs. Ontario in 2007 (cancer).

and private care. Due to its potential conflicts with the Canada Health Act, some people believe that this ruling might lead to the dismantling of Medicare, while others suggest that this is a strong call for reforming the current single-payer health system. Today long waiting times and privatization are among the most controversial topics in the health care reform debates in Canada and other OECD countries.[2] Our study is motivated to investigate whether or not allowing private care financing is associated with shorter waiting times in a public health system.

To answer the question, we employ a unique data set which encompasses joint replacement surgery records of nine Canadian provinces. The information contained in this data set has been used in Canadian Joint Replacement Registry (CJRR) annual reports (e.g., Canadian Institute for Health Information 2008a), but it is the first time that these data are examined with econometric methodologies. To our best knowledge, CJRR is the only nationwide data registry that collects waiting time information from different provinces for a specific surgical procedure using the same methodologies. Our main hypothesis for empirical tests is developed through the comparative statics of a demand and supply equilibrium. A province's public health system is assumed to be a general queuing system, so observations are constructed at the province/joint/month level to reflect the statistics of the queuing system. Lags of arrival rates are proposed to capture the "lagging effects" of arrival rates on waiting times. We employ both the generalized least square model and Random-effects model. We find that policies that induce private care financing appear to be associated with shorter public waiting times. The results also suggest positive lagging effects of arrival rates on waiting time. Sensitivity analysis is conducted to check the robustness of the findings.

The existing literature on private health care mainly focuses on two dimensions. One dimension is related to the financing of private care (e.g., the discussions of private insurance, tax incentives to private insurance and cost-sharing), and the other dimension is related to the ownership of care provision (e.g., the discussion of physician's private practice). Our study focuses on the dimension of private care financing.

[2] For instance, in the health reform debates of the United States in 2009, Canada's waiting time problem was described by many people as a seemingly unavoidable issue associated with universal health care (e.g. Wendell Goler, *Canada's Health System Informs U.S. Health Care Debate*, http://www.foxnews.com/politics/2009/08/12/ (August 12, 2009)). In particular, the problem was used to counter the idea of introducing a public, government-administered health insurance to compete with private insurance (Blendon and Benson, 2009).

In this regard, this study distinguishes from other studies that focus on private care provision or focus on both dimensions (e.g., Iversen 1997). As suggested by our institutional context, this study assumes that all health care services are delivered privately, so the difference between public and private care rests only on the financing side. Accordingly, in the rest of this chapter, “public care” or “public health service” refers to services that are funded through social insurance programs, while “private care” or “private health service” refers to services that are funded through private sources.

Our study distinguishes from the previous empirical studies of private care financing in the following ways. First, using case studies and simple statistical analysis, Tuohy et al. (2004) and Siciliani and Hurst (2005) compare aggregate data from different country models in which demographic characteristics and regulatory environments are not controlled for. By contrast, the control groups in our study are more homogenous with respect to these factors, except for the policies concerning private care financing, so our study is in a better position to tease out the effects of private care financing on the public waiting time. Second, physician dual practice is not allowed in our setting, which minimizes the potential concern of physician’s incentive distortions. For instance, if a physician is allowed to provide both public and private care, he may strategically manipulate the public waiting list so as to increase the demand for private care. Third, the use of econometric techniques distinguishes our study from other studies that utilize Canadian data. For instance, Cipriano et al. (2007) use Ontario joint replacement data to develop simulation models for waiting time prediction and policy evaluation. To our best knowledge, Besley et al. (1999) and Jofre-Bonet (2000) are the only studies that employ econometric models. However, their perspectives are different from ours: both studies consider the long public waiting list as a driver to the demand for supplemental private health insurance. Finally, our study contributes to Canada’s health care reform debates. Most of the current debates may be characterized much more by claims and counterclaims from competing ideological bases than by evidence (Tuohy et al. 2004). Rigorous empirical investigations of Canadian evidence are relatively rare, which is due partly to lack of data collection and partly to the complicated nature of health waiting times. For instance, there is no consensus about the definition and measurement of “medically necessary” waiting times. In this sense, it presents us a unique opportunity to answer the question as which claims can be substantiated.

The remainder of this chapter is organized as follows. Section 2.2 describes Canadian institutional background. Section 2.3 develops a theoretical model of a demand and supply equilibrium in the public health system, and derives main hypotheses for empirical testing. Section 2.4 describes the data, sets up the econometric specifications and discusses the relevant econometric issues. Section 2.5 presents the empirical results, and Section 2.6 concludes.

2.2 Institutional Background

Canada's public health system is a single-payer universal health system. Medically necessary health services are funded through general taxation and are provided by Medicare free of charge at the point of consumption. Evans (2000) provides an overview of the structure and funding of Canada's health system. Physicians in Canada are considered as for-profit independent entrepreneurs. Nevertheless, physicians need to choose between wholly staying in the public system (opt-in) and wholly staying out of the public system (opt-out). Opt-in physicians are eligible to bill the social insurance plan whilst opt-out physicians do not get any payments from the social insurance plan (status disincentive). Canada Health Act does not directly prohibit privately funded medically necessary services, but many provinces do not allow opt-out physicians to extra-bill private patients and most provinces do not provide public subsidy to private care. The ban on the purchase of private insurance also limits patients' choice of private care (Flood and Archibald, 2001). Canada's prohibition of physician dual practice is also in contrast to the general practice in other OECD countries. Furthermore, all Canadian provincial health systems need to meet the five guidelines of Canada Health Act to receive federal transfers, but provinces do regulate private care financing and physician's private practice in different ways. Flood and Archibald (2001) provide a review of these regulatory differences.

Long public health care waiting time is at the heart of the health reform debates in Canada. The call for privately funded health care is heightened by lawsuits such as the famous Chaoulli vs. Quebec case in 2005. Despite the measures that federal and provincial governments have taken to tackle long waiting times, some people believe that the waiting time problem roots in Canada's single-payer health system – patients do not have a choice in face of long waiting times. Supporters of this rationale propose to allow patients to purchase privately funded health care. The argument is that part of the demand for public care will be shifted to the private sector so that congestion in the public health system will be mitigated. Furthermore, the private sector is likely more efficient in utilizing resources than the public sector. Nevertheless, some researchers argue that the introduction of a two-tier system will not reduce the long public waiting time, if not worsening it.[3]

Although private care is largely banned in Canada, there are examples of private clinics. For instance, private clinics of orthopedic surgeries include, but not limited to, Duval orthopedic clinic, Westmount square surgical center, and RocklandMD surgery center in Quebec, Cambie surgery center, False Creek healthcare center and Kamloops surgical center in British Columbia, and Maples surgical center in Manitoba. In effect, no formal survey or study has been conducted to investigate the private sector in Canada and the extent of breach of Canada Health Act.

2.3 Theoretical Model and Empirical Implication

2.3.1 The Demand-Supply Equilibrium

In this section, we shall consider a demand and supply equilibrium for public care and then derive the hypothesis for empirical test. Let λ be the demand (patient arrival rate) for public care, w be the average public waiting time and η be the extent of privately funded care, then λ may be written as a function of w and η:

$$\lambda = \lambda(w, \eta) \tag{2.1}$$

The demand function is assumed to have the following properties. First, $\partial\lambda/\partial w < 0$: when the public waiting time increases, less patients are willing to join the public health system. Second, $\partial\lambda/\partial\eta < 0$: when the extent of privately funded care increases, the supply of private care will increase and the market for private care will grow, so more patients are willing to choose private care, moving away from the

[3] Evans (2000) argues that a truly private, parallel system of care, which is self-financing and independent of the public system, will allow health providers to charge patients fees in addition to the negotiated fee schedules between the provincial governments and the medical associations in return for perceived preferred access to care, while remaining fully eligible to bill the public system. He contends that the biggest risk of a two-tier system is that practitioners would be able to manipulate patient's access to public facilities and services in various ways so as to induce or compel patients to pay extra "private" fees, which would result in a greater social inequality.

public system.

Similar to Martin and Smith (1999), the supply function of public care is modeled through the optimization problem of the key decision maker of the public health system, e.g., the provincial health ministry. The production of public health services is constrained by two types of resources: human resources and non-human resources. Human resources include nurses, physicians and other hospital personnel that work in the public health system. Non-human resources are physical assets such as hospital facilities. Human resources and non-human resources are functions of η, denoted by $B_{HR}(\eta)$ and $B_{NHR}(\eta)$ respectively. We assume $\partial B_{HR}/\partial\eta \leq 0$, i.e., if the extent of privately funded care increases, the market for private care will grow so that some human resources will be attracted to the private sector. We assume $\partial B_{NHR}/\partial\eta = 0$, i.e. the extent of privately funded care has no impact on the physical assets since private sector can invest in these resources using private payments from the patients.

The health ministry needs to allocate $B_{HR}(\eta)$ and $B_{NHR}(\eta)$ to various types of public health services. Some of the health services (e.g., community care, acute care and long term care) are not prone to privatization so they can only be funded through general taxation. Let N be the supply of public health services that are not prone to privatization, and $v(N)$ be the amount of health surplus generated from these health services. We assume $\mathrm{d}v/\mathrm{d}N > 0$ and $\mathrm{d}^2v/\mathrm{d}N^2 \leq 0$. Let μ be the supply of public health services that are prone to privatization (which we call as "public care" in the rest of this section). We assume that each unit of public health service costs one unit of human resources and one unit of non-human resources.

According to queuing theories, waiting time is a function of the arrival (demand) rate and the service (supply) rate, denoted as μ, so we have a general waiting time function as

$$w = f(\lambda,\mu) \tag{2.2}$$

which has direct effects $\partial f/\partial\lambda > 0$ and $\partial f/\partial\mu < 0$. We further assume $\partial^2 f/\partial\mu\partial\lambda \leq 0$ and $\partial^2 f/\partial\mu^2 \geq 0$.[4]

Let R be the health benefit that a patient obtains from public care, θ be the cost of one unit of time in waiting, then λR is the total amount of health benefits that patients obtain from the public system and $\theta\lambda f(\lambda,\mu)$ is the total cost of waiting. We

[4] For instance, these signs hold for the queuing time of $M/M/1$ queue.

assume that the health ministry's objective is to maximize the total amount of health surplus out of resource $B_{HR}(\eta)$ and $B_{NHR}(\eta)$ given a demand rate of λ.[5] Therefore, the health ministry's optimization problem is:

$$
\begin{aligned}
\max_{\mu} \quad U(\mu) = v(N) \quad &+ \quad \lambda R - \theta\lambda f(\lambda,\mu) \\
\text{s.t. } N \quad &+ \quad \mu \le B_{HR}(\eta) \qquad (2.3) \\
N \quad &+ \quad \mu \le B_{NHR}(\eta) \qquad (2.4)
\end{aligned}
$$

Since the objective function is increasing in N and μ, one of the two constraints, (2.3) and (2.4), must be binding. Let $B(\eta) = \min\{B_{HR}(\eta), B_{NHR}(\eta)\}$ be the binding constraint, then the optimal supply, μ^*, is determined by the first order condition as:

$$\frac{\mathrm{d}v}{\mathrm{d}N} + \theta\lambda \frac{\partial f}{\partial \mu} = 0 \qquad (2.5)$$

Therefore, μ^* is a function of λ and η, i.e.,

$$\mu^* = \mu(\lambda, B(\eta)) \qquad (2.6)$$

We take the partial derivatives with respect to λ and η at both sides of (2.5) and reshuffle the terms, then we have direct effects $\partial\mu/\partial\lambda$ and $\partial\mu/\partial\eta$ as follows:

$$
\begin{aligned}
\frac{\partial \mu}{\partial \lambda} &= -\left(\theta\lambda \frac{\partial^2 f}{\partial\mu\partial\lambda} + \theta \frac{\partial f}{\partial \mu}\right) / \left(\theta\lambda \frac{\partial^2 f}{\partial \mu^2} - \frac{\mathrm{d}^2 v}{\mathrm{d}N^2}\right) > 0 \qquad (2.7) \\
\frac{\partial \mu}{\partial \eta} &= -\frac{\mathrm{d}^2 v}{\mathrm{d}N^2} \frac{\partial B}{\partial \eta} / \left(\theta\lambda \frac{\partial^2 f}{\partial \mu^2} - \frac{\mathrm{d}^2 v}{\mathrm{d}N^2}\right) \le 0 \qquad (2.8)
\end{aligned}
$$

Given the partial derivatives of $f(\lambda,\mu)$ and $v(N)$, it is straightforward to see that $\left(\theta\lambda \cdot \partial^2 f/\partial\mu^2 - \mathrm{d}^2 v/\mathrm{d}N^2\right) > 0$ and $\left(\theta\lambda \cdot \partial^2 f/\partial\mu\partial\lambda + \theta \cdot \partial f/\partial\mu\right) < 0$, so $\partial\mu/\partial\lambda > 0$. This means that if the demand for public care increases, the health ministry has an incentive to allocate more resources to the production of public care. For instance, Cipriano et al. (2007) discuss the minimum growth rates of supply of joint

[5] In our setting, it is reasonable to assume that λ is given in the health ministry's decision making. This is because the health ministry does the planning based on demand forecast for a relatively long time horizon, e.g., one year. Once the planning decisions N and μ are made, they are unlikely to be affected by short term demand changes.

replacement surgeries in Ontario in response to a wide range of demand projections over a 10-year period in order to achieve waiting time targets. As for inequality (2.8), if $B(\eta) = B_{HR}(\eta)$, then $\partial B/\partial\eta = \partial B_{HR}/\partial\eta \leq 0$. Therefore, we have $\mathrm{d}^2v/\mathrm{d}N^2 \cdot \partial B/\partial\eta \geq 0$ and thus $\partial\mu/\partial\eta \leq 0$. This means that some human resources in the public health system are drawn to the private sector so the supply of public care is reduced. If $B(\eta) = B_{NHR}(\eta)$, then $\partial B/\partial\eta = \partial B_{NHR}/\partial\eta = 0$ and thus $\partial\mu/\partial\eta = 0$. In this case, the existence of private sector has no impact on the supply of public care.

The demand-supply equilibrium should simultaneously satisfy equation (2.1), (2.2) and (2.6). Inputting (2.1), (2.6) into (2.2) yields the waiting time equation as:

$$w = f(\lambda(w,\eta), \mu(\lambda(w,\eta),\eta)) \tag{2.9}$$

We assume the existence of the equilibrium in the same way as Linsay and Feigenbaum (1984). Solving (2.9) for w gives rise to the equilibrium waiting time as a function of η:

$$w = h(\eta) \tag{2.10}$$

2.3.2 Effects of Policies 1 and 2 on Public Waiting Time

Two policies are available for our empirical tests. Policy 1 allows opt-out physicians to extra-bill private patients, so more health service providers are willing to participate in the private care market and the supply of private care increases. Policy 2 provides public subsidies to patients seeking private care, so more patients are willing to choose private care and the demand for private care increases. Therefore, both policies increase the extent of privately funded care. The effects of policy 1 and policy 2 on the public waiting time are obtained through the net effect of η on w. We derive the net effect of η on w through comparative statics. Taking derivative with respect to η at both sides of equation (2.9) yields:

$$\frac{\mathrm{d}w}{\mathrm{d}\eta} = \frac{\partial f}{\partial\lambda}\left(\frac{\partial\lambda}{\partial w}\frac{\mathrm{d}w}{\mathrm{d}\eta} + \frac{\partial\lambda}{\partial\eta}\right) + \frac{\partial f}{\partial\mu}\left(\frac{\partial\mu}{\partial\lambda}\left(\frac{\partial\lambda}{\partial w}\frac{\mathrm{d}w}{\mathrm{d}\eta} + \frac{\partial\lambda}{\partial\eta}\right) + \frac{\partial\mu}{\partial\eta}\right)$$

which must lead to:

$$\frac{\mathrm{d}w}{\mathrm{d}\eta} = \left\{ \frac{\partial f}{\partial \lambda}\frac{\partial \lambda}{\partial \eta} + \frac{\partial f}{\partial \mu}\left(\frac{\partial \mu}{\partial \lambda}\frac{\partial \lambda}{\partial \eta} + \frac{\partial \mu}{\partial \eta}\right)\right\} / \left(1 - \left(\frac{\partial f}{\partial \lambda} + \frac{\partial f}{\partial \mu}\frac{\partial \mu}{\partial \lambda}\right)\frac{\partial \lambda}{\partial w}\right) \tag{2.11}$$

The numerator of the right hand side of (2.11) has two competing effects:

$$\frac{\partial f}{\partial \lambda}\frac{\partial \lambda}{\partial \eta} < 0 \tag{2.12}$$

$$\frac{\partial f}{\partial \mu}\left(\frac{\partial \mu}{\partial \lambda}\frac{\partial \lambda}{\partial \eta} + \frac{\partial \mu}{\partial \eta}\right) > 0 \tag{2.13}$$

We call (2.12) the demand-side effect and (2.13) the supply-side effect. A negative demand-side effect means that a higher extent of privately funded care induces patients to switch from public care to private care. When the non-human resources are the bottleneck of the public health system, we have $\partial\mu/\partial\eta = 0$. In this case, a positive supply effect means that as the demand for public care decreases, the health ministry allocates less resources to the production of public care. When the human resources are the bottleneck of the public health system, a positive supply-side effect means that apart from less budget allocation, the increase of privately funded care "crowds out" the human resources used for the production of public care, which increases the public waiting time.

If the demand-side effect dominates the supply-side effect, we must have

$$\begin{aligned} & \frac{\partial f}{\partial \lambda}\frac{\partial \lambda}{\partial \eta} + \frac{\partial f}{\partial \mu}\left(\frac{\partial \mu}{\partial \lambda}\frac{\partial \lambda}{\partial \eta} + \frac{\partial \mu}{\partial \eta}\right) < 0 \\ \Rightarrow \quad & \frac{\partial f}{\partial \lambda} + \frac{\partial f}{\partial \mu}\frac{\partial \mu}{\partial \lambda} \; > \; -\frac{\partial f}{\partial \mu}\frac{\partial \mu}{\partial \eta} / \frac{\partial \lambda}{\partial \eta} \; \geq 0 \end{aligned} \tag{2.14}$$

Since $\partial\lambda/\partial w < 0$, by (2.14), the denominator of (2.11) is positive and it immediately yields

$$\frac{\mathrm{d}w}{\mathrm{d}\eta} < 0$$

On the other hand, if the supply-side effect dominates the demand-side effect, we must have

$$\begin{aligned} & \frac{\partial f}{\partial \lambda}\frac{\partial \lambda}{\partial \eta} + \frac{\partial f}{\partial \mu}\left(\frac{\partial \mu}{\partial \lambda}\frac{\partial \lambda}{\partial \eta} + \frac{\partial \mu}{\partial \eta}\right) \geq 0 \\ \Rightarrow \quad & \frac{\partial f}{\partial \lambda} + \frac{\partial f}{\partial \mu}\frac{\partial \mu}{\partial \lambda} \;\leq\; -\frac{\partial f}{\partial \mu}\frac{\partial \mu}{\partial \eta}\Big/\frac{\partial \lambda}{\partial \eta} \end{aligned}$$

In this case, the sign of $\left(\frac{\partial f}{\partial \lambda} + \frac{\partial f}{\partial \mu}\frac{\partial \mu}{\partial \lambda}\right)$ is undetermined, which makes the sign of the denominator of (2.11) and further the sign of $\mathrm{d}w/\mathrm{d}\eta$ undetermined. In summary, if the demand-side effect dominates the supply-side effect, an increase of the extent of private care would result in a longer public waiting time; If the supply-side effect dominates the demand-side effect, the impact of an increase of the extent of private care on public waiting time is undetermined.

Our simple theoretical model thus suggests that the effects of policy 1 and policy 2 on the public waiting time depend on the relative strengths of the competing demand-side and supply-side effects. The negative demand-side effect corresponds to the well-known supportive argument for private care – the existence of private care lessens the burden of the public health system. In discussing the competing explanations of public waiting lists, Cullis and Jones (1985) argue that providing subsidies to private care is more efficient to reduce the public waiting time than increasing the budget of public care. Iversen (1997) finds that the net effect of a private sector on the public waiting time depends on the relative strengths of the competing capacity and demand side effects. In his model, the strengths of the two competing effects are determined by the elasticity of the demand for public care with respect to the public waiting time. The more elastic the demand is, the more likely the existence of a private sector will increase the public waiting time. However, Iversen (1997) only models the case when the non-human resources are the bottleneck of the public health system. In contrast, our model also considers the case when the human resources are the bottleneck of the public health system, in which case the supply-side effect is determined by not only the health ministry's supply decision (i.e. $\partial f/\partial \mu \cdot \partial \mu/\partial \lambda \cdot \partial \lambda/\partial \eta$) but also the extent of privately funded care (i.e. $\partial f/\partial \mu \cdot \partial \mu/\partial \eta$).

On the empirical side, Siciliani and Hurst (2005) argue that the policies governments use to reduce the public waiting time can be categorized into supply side policies (i.e., inducing more supply of public care), demand side policies (i.e., reducing demand for public care) and policies acting directly on waiting times. Using evidence from Australia, the authors show that providing tax incentives to the purchase of supplemental private health insurance results in a dominant demand side effect, thereby reducing public waiting times (p.211 of Siciliani and Hurst (2005)). The authors suggest that governments should provide tax incentives for patients to purchase private health insurance and encourage patients to substitute private care for public care. In accordance with these theoretical findings and empirical evidence, we have the following main hypothesis for empirical tests:

Hypothesis 2.1 *Both policy 1 and policy 2 are associated with shorter public waiting times.*

2.4 Data and Econometric Models

We now describe, in Section 2.4.1, the data used for our empirical tests and construct the dependent and main explanatory variables. Additional explanatory variables suggested by the existing literature will be introduced. We then in Section 2.4.2 present the feasible generalized least square model and the random-effects model, as well as discuss the econometric issues related to lag variables and the random-effects specification.

2.4.1 Data and Variable Construct

The data for our econometric analysis are from Canadian Joint Replacement Registry (CJRR) which is administered by Canadian Institute for Health Information (CIHI). To our best knowledge, CJRR is the only nationwide data registry that collects waiting time information from different provinces for a specific surgical procedure using a consistent methodology. To comply with CIHI's strict standards of data release and security, our data application took one year to complete. Research ethics approvals and data security agreements at both the organizational and provincial levels are required for the final release of the data.

The actual data for empirical tests are comprised of 29,369 joint replacement surgery records of nine Canadian provinces based on admission date from April 1, 2005 to March 31, 2007. Each record contains a patient's demographic information (age, gender, distance from home to hospital), surgical information[6] (hip/knee,

[6] Other surgical information is excluded from our analysis due to a large percentage of data missing.

primary surgery/revision), waiting time information (the calendar month of making decision for surgery, the calendar month of admission to hospital,[7] and waiting time defined as the number of days from decision date to admission date). Since revision surgeries are to repair the primary joint replacements, we only include primary surgery data for analysis.[8]

Canadian orthopedic surgeons submit their joint replacement surgery records to CJRR on a voluntary basis. Depending on a province's surgeon participation rate and data submission,[9] the percentages of surgeries included in our data set (we call it "data inclusion" hereafter) varies by province, ranging from 3.6% of hip/knee replacements of Ontario in 2005/6 to 70% of knee replacements of New Brunswick in 2006/7.[10] To account for the voluntary nature of the data set, it behooves us to access the representativeness of our data.

First, we compare our waiting time data with the waiting time data published in Fraser Health Institute's (FHI) annual waiting time reports of 2005 and 2006 (Esmail and Walker, 2005, 2006), and we find a strong consistency.[11] Secondly, we compare the patient demographics of our data[12] with the patient demographics of Hospital Morbidity Database (HMDB) and Discharge Abstract Database (DAD), both of which contain all joint replacement hospitalizations during the same period. The patient demographics of all sources are very similar, if not closely matched.[13] To deal with outliers of waiting time and distance data, we follow the common procedure of winsorizing the data - we replace the top (bottom) 1% of the data by the 99

[7] CJRR only agrees to provide decision dates and admission dates in the form of calendar month.

[8] A total of 2,277 revision records, together with 51 erroneous primary records which have decision date after admission date, are removed.

[9] Canadian Institute for Health Information (2008a) estimates that all of New Brunswick and Nova Scotia orthopedic surgeons are participating in CJRR while only 44% of Quebec orthopedic surgeons are doing so. However, the actual surgeon participation rates are unknown. Also, a participating surgeon may not submit all of his patient records to CJRR.

[10] The detailed information of data inclusion is shown in Table A.1 of Appendix A.

[11] We compare the median waiting times (in province/year level; hip and knee combined) of our data with the median waiting times of arthroplasty surgeries (in province/year level; hip, knee, ankle and shoulder surgeries combined) from FHI's annual waiting time reports of 2005 and 2006. The two sets of median waiting times are strongly positively correlated - the correlation is 0.78 for the year of 2005 and 0.87 for the year of 2006. FHI's annual reports are based on nationwide survey data collected by the institute. To our best knowledge, FHI's annual reports are the only source that contains waiting time comparisons in the specialty level and uses data collected by consistent methodologies.

[12] Age, gender and joint type distributions of the 29,369 records.

[13] Our preliminary analysis also shows that patient demographics are similar among provinces.

(1) percentile value for each province-joint pair.[14]

The data are structured to reflect a province's monthly queuing statistics. We define province-joint pair as the cross section unit, so observations are at the province/joint/month level of analysis. In the rest of this chapter, we use subscript i to denote cross section unit and subscript t to denote calendar month.

The first control variable to be included in the regression model is the patient arrival rate, denoted by λ_t, which is measured as the count of patient records by decision date t. Accordingly, the dependent variable, w_t, is measured as the mean waiting time of these λ_t patients – prospective waiting time. Arrival rates and waiting times measured as above reflect the transient states of a queuing system. The transient waiting times are dependent on the history of arrival rates and service rates. Unfortunately, we are not able to include the time-dependent service rate as a control variable due to the unavailability of data. Nevertheless, since hospitals plan their resources on a regular basis for surgeries, e.g., 6 months in advance in British Columbia, we do not expect the service capacity to exhibit as much randomness as the arrival rates during the short time horizon of our study.

Prospective waiting times are consistent with our assumption in the demand function that patients make their decisions based on the expected waiting time. However, prospective waiting times come with the cost of data truncation.[15] Data truncation induces a bias towards the underestimation of λ_t and w_t, especially when the estimation moves closer to March 2007.[16] To ensure that the time series data of different cross section units in the analysis are comparable, for each cross section unit, we only include the data of those months of which at least 90% of the incoming patients are served by the end of the 24-month period.[17]

We expect arrival rates to have positive lagging effects on (prospective) waiting

[14] We also experiment with 2.5% and 5% cut-offs. The exact location of winsorization does not affect our results.

[15] Patients records with decision dates within the 24-month period but admission dates after March 2007 are not included in our data.

[16] On the other hand, Table A.1 of Appendix A shows that many provinces have a mild increase in data inclusion from 2005/6 to 2006/7, which may counteract the underestimation of λ_t due to data truncation.

[17] For instance, for cross section unit "Alberta-hip", 90% of its patient records have waiting times less than or equal to 11 months, i.e., at least 90% of the patients arriving in the first 13 months are served by the end of the 24-month period. Therefore, the first 13 months of data (April 2005 to April 2006) are used for "Alberta-hip". Table A.2 in Appendix A shows the number of months included in the analysis for each cross section unit, which ranges from 0 of Saskatchewan (knee) to 18 of Newfoundland (hip/knee).

times;[18] hence we propose lag variables λ_{t-j}, $j = 1, 2, ..$ in the econometric models. We use models with different number of lags to test the lagging effects and check the robustness of our main findings.[19] We normalize arrival rates by province-specific population and data inclusion.

The constructs of the main explanatory variables - indicator variables of policy 1 (i.e., whether opt-out physicians are allowed to extra bill private patients) and policy 2 (i.e., whether public subsidies are provided to private care) - use information from Flood and Archibald (2001). Moreover, in order to yield more robust causal inferences, it behooves us to control for other potential drivers of waiting times: namely, age, gender, joint type, distance and trending. Among them, the interaction effects of age and gender are examined. No other interaction effects are suggested by existing literature or found significant in the preliminary analysis of differences between means.

(i) Age and gender: Canadian Institute for Health Information (2008a) reports waiting time differences by gender and age, and age differences by joint type.[20] A possible explanation of the impact of age and gender on waiting times is that age and gender correlate with patient's obesity.[21] Obese patients may be requested to lose weight prior to undergoing surgeries to improve the outcome of the procedure (Canadian Institute for Health Information, 2008a). In the econometric models, we use variable "femaleage" and "maleage" to isolate the age effect by gender, and use variable "female" (the percentage of female patients) to capture the gender effect.

(ii) Hip/knee: Canadian Institute for Health Information (2008a) finds that in general hip patients have shorter waiting times than knee patients. However, the report does not provide elaboration about the cause of this difference.

[18] If we consider each surgeon as a single server queue, then the elementary queueing theory implies that the arrival rates have positive lagging effects on the waiting times. We expect such positive lagging effects to prevail in a system with many surgeons in aggregation.

[19] Our estimation of the arrival rates prior to April 2005 is as follows: for month t prior to April 2005, we estimate that $y\%$ of patient records are not included in our data. Let c be the count of surgery records for month t and thus we estimate $\lambda_t = \frac{c}{1-y\%}$.

[20] Canadian Institute for Health Information (2008a) finds that female patients tend to have shorter waiting times than male patients. Knee patients are significantly older than hip patients.

[21] Karlson et al. (2003) find that higher Body Mass Index and older ages significantly increase the risks of osteoarthritis, the most commonly reported diagnosis of joint replacement patients.

(iii) Distance: our preliminary data analysis shows that patients in provinces of Manitoba, Newfoundland, Nova Scotia and Saskatchewan tend to have longer distances from home to hospital than patients in other provinces. Patients in these provinces may have extra difficulties to access health services as these provinces have low population densities and lack of major urban areas where health care facilities are located close to the majority of population.

(iv) Trending: our preliminary analysis of the mean waiting times by admission date shows a declining time trend for some cross section units.[22] Provincial governments' short term measures to deal with patient backlog may contribute to this time trend.[23] Also, the data truncation problem (as discussed above) is likely to result in a declining trend in the estimation of λ_t and w_t. When both the dependent and explanatory variables are trending, Wooldridge (2009) suggests adding a time trend variable to obtain a detrending interpretation of the regression. Therefore, we include a time trend variable, "month", which measures the number of months after March 2005, in the econometric model.[24]

Another possible explanatory variable that might be considered is the *ratio of general physicians to orthopedic surgeons*. General physicians (upstream) and orthopedic surgeons (downstream) are at different stages of a patient's care path, so the public health system can be seen as a tandem queue. With everything else being equal, congestion (e.g., long waiting time) more likely appears at a stage where service capacity (e.g., number of service providers) is relatively scarce. However, preliminary analysis (see Table A.3 and Table A.4 in Appendix A) does not show an anticipated correlation between this explanatory variable and the dependent variable. Moreover, inclusion of this variable into regressions generates inconsistent signs of coefficient and insignificant results for this variable. Therefore, we decide not to report regression results with this variable being included.

Table 2.1 provides the source of variables and summary statistics. In consistent with the discussions above, we hypothesize that arrival rates have positive lagging

[22] There is a slight declining trend for Alberta (hip/knee), British Columbia (hip/knee) and Manitoba (hip/knee).

[23] Provincial governments were under pressure to achieve meaningful waiting time reductions for joint replacement surgeries by March 2007, so they had an incentive to increase the service provision; see the numbers of hospitalizations of 2005/6 and 2006/7 in Table A.1 of Appendix A.

[24] We also test models using indicator variables for calendar months, which does not change our results.

effects on the public waiting time, and that older patients, male patients, knee patients and patients having longer home-to-hospital distances have longer waiting times.

Table 2.1: Variables included in the regression models and descriptive statistics

Variable name	Definition	Source	Mean	Std dev	Min	Max
w	Mean waiting time (days)	CJRR	164.02	63.06	63.46	338.26
Month	Number of months after the starting month (April 2005)	CJRR	6.69	4.14	1	18
Hip	Indicator of hip replacement	CJRR	0.54	0.49	0	1
Distance	Mean distance of home to hospital (in kilometers)	CJRR	39.47	27.48	13.78	304.16
Policy1	Indicator if opt-out physicians are allowed to extra bill private patients	Flood and Archibald (2001)	0.719	0.45	0	1
Policy2	Indicator if public subsidy is provided to private care	Flood and Archibald (2001)	0.18	0.38	0	1
Female	% of female patients	CJRR	0.58	0.09	0.28	0.93
Femaleage	Mean age of female patients	CJRR	67.22	2.98	56.4	74.54
Maleage	Mean age of male patients	CJRR	65.17	4.37	41.75	75.67
λ	Arrival rate (number of incoming patients per 10^5 population) adjusted for province data inclusion	CJRR	8.59	2.94	2.70	17.50

2.4.2 Econometric Issues and Sensitivity Analysis

By (2.10) and including the explanatory variables discussed in Section 2.4.1, we have the following regression models:

$$\begin{aligned} w_{it} =& \beta_0 + \sum_{j=0}^{k} \beta_{1,j}\, \lambda_{i,t-j} + \beta_2\, poliy1_i + \beta_3\, poliy2_i + \beta_4\, month_t + \beta_5\, hip_i + \beta_6\, distance_{it} \\ & + \beta_7\, female_{it} + \beta_8\, femaleage_{it} + \beta_9\, maleage_{it} + \varepsilon_{i,t}, \;\; k = 1,2,... \end{aligned} \tag{2.15}$$

It should be noted that as the main purpose of this study is not to search for a model that best fits our data, we do not utilize techniques like variable and model selection to optimize the inclusion of control variables and the form of regression models. However, in order to properly employ our data, a couple of econometric issues need to be considered. First, due to the non-variation of the key variables (policy 1 and policy 2), we are unable to fully employ panel data econometric techniques, i.e., the Fixed-effects model. Therefore, we resort to the feasible generalized linear square (FGLS) model.[25] Nevertheless, to account for the unobserved cross section specific effects and for the sake of robustness check, we further examine our data using the random-effects specification. The random-effects model is constructed by adding a cross section specific term u_i to the econometric model (2.15), assuming that explanatory variables are exogenous to u_i.

Due to the general multicollinearity problem between lag variables, the t-statistics of lag variables are often less significant. Wooldridge (2009) suggests using the Wald-test to test the joint significance of lag variables.

To check for the robustness of our findings, we conduct sensitivity analysis as follows. The results of sensitivity tests are consistent with the results presented in Table 2.2.

- To account for the likely high and uncontrolled clustering of dependent variables, we run regressions with standard errors clustered at the province-joint level.[26] Comparing to the results of Table 2.2, the only difference is that the

[25] For other studies that utilize panel methods to empirically investigate waiting time issues, see Martin and Smith (2003); Siciliani and Martin (2007).

[26] This is done by adding "vce(cluster)" option in the Stata command for FGLS and Random-effects models.

coefficients of “Policy 1” become insignificant in FGLS model with clustering.

- Autocorrelation is a concern when running regressions on time series data. We perform the Wooldridge (2002) test for autocorrelation in panel data to our data set and the test result does not reject the null hypothesis of no first order autocorrelation. To avoid the potential correlation of error terms between joint types of the same province, we also test by running the regressions solely on hip or knee data. The results show that the estimation using knee data generates more significant results than the estimation using hip data.

- To account for province-specific waiting time trends, we can replace a common time trend by 18 province-joint specific time trends in the regressions. The results show a declining trend for 5 province-joint pairs, an increasing time trend for 2 province-joint pairs and no time trend for other province-joint pairs. In Table 2.2, we only present results for regressions with a common time trend as we believe it better reflects the macro movement of waiting times over the 24-month period.

2.5 Results and Discussion

Table 2.2 shows the results of both the FGLS model and the Random-effects model. FGLS model uses panel-specific AR(1) autocorrelation structure and allows for heteroskedasticity, but assumes no cross-sectional correlation. Random-effects model clusters standard errors at province-joint level. For both specifications, the Wald-test suggests that only when the 3-month and 4-month lags of arrival rates are included, the coefficients of lag variables are jointly significant. Residual analysis shows no signs of heteroskedasticity or abnormality for each province-joint pair. As this study is subject to small sample size, we need to interpret them with caution.

In the following discussions, we compare our results to those of CJRR annual report (Canadian Institute for Health Information, 2008b) wherever possible. The findings of CJRR annual report are based on observations on descriptive statistics. Therefore, our regression results should be deemed more creditable as we have properly controlled for different factors.

The coefficients of policy 1 and policy 2 suggest that both policies are associated with shorter waiting times, which supports our main Hypothesis 2.1. The results of policy 2 are more robust than those of policy 1. Additionally, the magnitude of coefficients of policy 2 is greater than that of policy 1. Both findings suggest that policy 2 is associated with a greater waiting time reduction than policy 1. This is consistent with the conclusion drawn by Flood and Archibald (2001) that public funding to private care is the key in making private care an option for patients. Policy 2 subsidizes patient's private care up to the amount of public fee schedule, which gives patients a substantial incentive to opt out of the public health system and shifts the demand from public health system to private care market. This corresponds to a dominant demand-side effect in (2.11).

The coefficients of arrival rates and lags are mostly positive. The results of Wald-test suggest that when the 3-month and 4-month lags of arrival rates are included in the models, the coefficients of lag variables are jointly significant.[27] The coefficients of the 3-month and 4-month lag variables suggest that if there is one more patient arrival in the current month, patients arriving three or four months later should expect an extra delay of one to three days over the average waiting time.

The coefficients of "Month" are negative and significant for all model specifications, suggesting a declining time trend. Data truncation problem may contribute to this time trend. More importantly, in September 2004 federal and provincial governments set up the "Wait Times Reduction Fund" (in *the 10-Year Plan to Strengthen Health Care*) to build capacities to address waiting time problems. Joint replacement surgery is identified as a target area for this fund. These government-led initiatives may affect the joint replacement surgery waiting times through the following channels. First, the fund could have directly increased the provision of joint replacement surgeries. Second, under the pressure of achieving meaningful waiting time reductions by March 2007, provincial governments might have relocated some capacities that were previously used by other services to joint replacement surgeries. Johnson et al. (2007) show that between 2001/2 and 2005/6, there was a dramatic increase of age-standardized rate for joint replacement surgeries, but a decrease of age-standardized rate for other surgeries.

The coefficients of "Hip" are negative for all model specifications but are only significant for FGLS model with 3-month lags of arrival rate. The results suggest that knee patients generally have to wait one to three weeks more than hip patients. This

[27] We also use F-test to test the model specifications - comparing the log likelihood of models with and without lag variables. Similar results are obtained.

finding is consistent with CJRR annual report (Canadian Institute for Health Information 2008a).

The coefficients of “Distance” are positive and significant for all model specifications. The results suggest that patients with longer distances have longer waiting times. The coefficients of “Distance” in the random-effects model are much smaller than those in the FGLS model, which suggests that distance impacts on patient’s waiting time partly through the province specific mean. This finding is consistent with our preliminary analysis of distance data.

In addition to the robust results discussed above, the coefficients of “Female” are negative but insignificant for all model specifications. The negative sign is consistent with the findings of Canadian Institute for Health Information (2008a) – female patients tend to have shorter waiting times than their male counterparts. This may be due to the higher obesity of male patients. As obesity positively correlates to poor health conditions, obese patients may require more medical treatments before surgeries. The regression results exhibit mixed age effects on waiting times. The coefficients of “Femaleage” are negative but insignificant for all model specifications, while the coefficients of “Maleage” are inconclusive.

2.6 Concluding Remarks

The objective of this study is to test the relationship between privately funded health care and the public waiting time. How private care impacts on the public health waiting time has been extensively debated in Canada and other OECD countries, but often debates are based on ideological arguments and claims rather than empirical evidence. Using a sample of joint replacement surgery data from Canada, we had investigated the relationship between policies that induce private care financing and the public waiting time. Based on the results of the FGLS model and the Random-effects model, we found that the policies of interest are associated with shorter public waiting times. According to the theoretical model in Section 2.3, the shorter public waiting times observed in empirical data can be attributed to a dominant demand-side effect introduced by the policies of interest, i.e., these policies do not reduce the provision of public care very much but induce patients to seek private care so that the demand pressure on the public health system is mitigated. Our results also suggested positive lagging effects of arrival rate on the public waiting time.

Although this study provides supportive evidence to one side of the public versus private debates, policy implications should be drawn with cautions as our study is subject to data limitation. In this regard, our study would motivate more empirical studies of public waiting time and a mixed financed health system. We look forward to applying the same methodologies to more recent data to investigate policy changes after the ruling of Chaoulli vs. Quebec took effect in January 2008.

Table 2.2: Regression results[a]

	FGLS model		Random-effects model	
Intercept	276.2	248.2	211.0	199.9
	(62.82)***	(63.07)***	(67.11)**	(75.64)**
Month	-2.404	-2.928	-1.917	-2.258
	(0.603)***	(0.587)***	(0.689)**	(0.750)**
Hip	-14.50	-7.703	-19.71	-14.38
	(6.842)*	(6.830)	(24.25)	(25.04)
Distance	0.559	0.490	0.301	0.313
	(0.123)***	(0.121)***	(0.150)*	(0.151)*
Policy 1	-38.06	-32.87	-25.86	-23.22
	(13.53)**	(13.72)*	(34.22)	(34.72)
Policy 2	-85.01	-78.94	-81.17	-77.63
	(7.938)***	(7.999)***	(11.86)***	(11.02) ***
Female	-16.07	-21.39	-24.46	-30.81
	(18.35)	(17.88)	(20.12)	(22.17)
Maleage	-0.317	-0.0335	0.250	0.261
	(0.483)	(0.486)	(0.585)	(0.667)
Femaleage	-0.976	-1.198	-0.115	-0.200
	(0.665)	(0.669)	(0.802)	(0.802)
λ_{-1}	0.968	1.442	0.319	0.448
	(0.678)	(0.672)*	(1.196)	(1.190)
λ_{-2}	0.742	0.546	-1.243	-0.872
	(0.661)	(0.659)	(1.117)	(1.209)
λ_{-3}	3.129	3.300	2.422	2.361
	(0.662)***	(0.655)***	(0.834)**	(0.790)**
λ_{-4}		2.462		1.454
		(0.671)***		(1.399)
Observations	184	184	184	184
Log likelihood	-827.38	-826.74		
R-squared			0.63	0.67
Wald-test that coefficients of λ and lag variables are all equal to zero (i.e., Prob > chi2)	0.000	0.000	0.008	0.009

[a] ***, ** and * denote statistical significance at the 1%, 5% and 10% confidence levels, respectively. Standard errors are shown in parentheses.

Chapter 3

Physician Dual Practice, Public Waiting Time and Patient Welfare

3.1 Introduction

This study is motivated by cataract surgery evidence in the province of Manitoba, Canada (DeCoster et al., 2000). Before January 1999, cataract patients of Manitoba needed to pay a "facility" fee of $1,000 to obtain surgeries in privately-run clinics, while they paid no fees for surgeries in the public hospitals. Ophthalmological surgeons were allowed to practise in both public hospitals and private clinics. These surgeons are referred as "dual-practice physicians". On the other hand, there were surgeons who only practised in the public hospitals and thus are referred as "public-only physicians". The provincial social health insurance program paid the same amount of fee-for-service to each cataract surgery regardless where it was supplied. In this case, the social insurance subsidized private surgeries up to the amount of public fee schedule. After January 1999, the facility fee of $1,000 (referred as "extra billing") was banned. However, the provincial social insurance agreed to pay for the same number (but no more) of surgeries in the private clinics in the last year prior to the ban. Over time the private clinics were merged into the public system. Today all cataract surgeries offered in Manitoba are considered as public surgeries and there are no more dual practice physicians.

During the period when extra billing was allowed, DeCoster et al. (2000) show in Table 3.1 that the waiting times of surgeries differ by physician practice type and the location of provision.

Table 3.1: Median waiting times (in weeks) of cataract surgeries in Manitoba: 1992-1999[a]

	Year	92/93	93/94	94/95	95/96	96/97	97/98	98/99
Public-only physician	Public hospital	14	8	6	6	10	10	10
Dual-practice physician	Public hospital	18	14	14	19	23	21	26
	Private clinic	4	4	4	4	4	5	5

[a] Source: DeCoster et al. (2000).

The first row of Table 3.1 is the median waiting times of surgeries performed by public-only physicians in public hospitals. The second and third rows are the median waiting times of surgeries performed by the same dual-practice physicians in public hospitals and private clinics, respectively. Two types of waiting time difference can be observed in Table 3.1. First, the last two rows of Table 3.1 show that the waiting times of private surgeries are far shorter than the waiting times of public surgeries, all of which were performed by the same dual-practice physicians. This waiting time difference may be attributed to the priority that dual-practice physicians gave to private patients. A dual-practice physician treated all incoming patients equally until the point when the decision to surgery was finalized and patients were offered the option of private surgery. If a patient opted for private surgery by paying the extra facility fee, the dual-practice physician would arrange the surgery in a private clinic at the earliest possible date. Therefore, it is reasonable to assume that some priority is at work for private patients in the queues of dual-practice physicians. Second, the first two rows of Table 3.1 show that the waiting times of public surgeries provided by public-only physicians are shorter than the waiting times of public surgeries provided by dual-practice physicians. Neither price rationing nor physical environments seem to contribute to this waiting time difference, because no fees were charged for public surgeries and both types of physicians used the same public hospital facilities. Hypothetically, if the two types of physicians were identical in every aspect of service provision, and meanwhile the service attributes are observable to patients, then such a waiting time difference should not exist.

The above observations motivate us to provide a possible explanation for the waiting time difference of public surgeries between different physician practice types.

The analysis is based on certain assumptions that are either directly implied in Manitoba's empirical evidence or supported by the existing literature. First, we assume that physicians may be differentiated by their service qualities. The concept of "quality" here is an abstract term, which aggregates all attributes of a service except for the waiting time. This assumption is supported by the pricing mechanism observed in Manitoba's empirical evidence. The existing literature also shows that allowing physician dual practice could improve dual-practice physician's service quality, in which service quality can be defined as the appropriate amount of treatment (Rickman and McGuire, 1999), the accuracy of diagnosis (González, 2004), or an output determined by physician skill level (Bir and Eggleston, 2003; Biglaiser and Ma, 2007). We assume that service qualities are observable to patients in the case when physician dual practice is allowed. The observability is fulfilled through, for instance, price discrimination and/or discovering physician's practice type. Second, the waiting time difference observed in the surgeries of dual-practice physicians implies that private patients were given some sort of priority over public patients when a dual-practice physician managed her waiting list. In reality, dual-practice physicians may prioritize patients by considering more factors. To keep our analysis to the core of the problem, we assume that a pure priority is at work. It would be of our interest to see whether a model based on these assumptions would yield results that are consistent with Manitoba's evidence.

Our study shows that the waiting time difference of public surgeries existing between dual-practice physicians and public-only physicians could be explained by service quality differentiation. Additionally, the model based on the above assumptions allows us to investigate the impact of allowing physician dual practice on patient's waiting time and welfare. Our study shows that patients in the public queue of dual-practice physicians have to endure a longer waiting time than they would do in the case when physician dual practice is not allowed. However, there are always patients in the public queue of dual-practice physicians who would be better off by allowing physician dual practice, as they would enjoy a higher service quality from dual-practice physicians.

The model used in our study follows general results of queuing theories. Patients are assumed to be heterogenous with respect to their time costs. Two sets of health providers are in the health care market, i.e. dual-practice physicians and public-only physicians. Each set of providers serve a common stream of incoming patients. The queue of dual-practice physicians is modeled as a single queue with two classes of patients,[1] while there is only one single class in the queue of public-only physicians. Therefore, there are three service alternatives: private care, public care of dual-practice physicians and public care of public-only physicians. An incoming patient has information of the expected waiting time of each service alternative. The patient would choose the service alternative that offers the maximum expected net benefit or choose outside services that offer a reservation benefit. The expected net benefit is defined as the service benefit net of the price (if any) and the expected waiting cost.

In Section 3.4 to Section 3.6, we limit our discussions to the case when no patients seek outside service. This is, for one thing, to simplify the analysis so that some structural results can be obtained. Moreover, empirical evidence has revealed that the cost of cataract surgery in another jurisdiction could be very high for Canadian patients if they choose to pay out of their pockets. For example, the straightforward cataract surgery alone in the United States would cost a patient about $3,279 per eye nowadays,[2] let alone other ancillary costs such as the costs of transportation and accommodation. Therefore, patients who can afford such an expensive outside service only constitute a small fraction of the whole patient population. In other words, ignoring outside service in our model would not undermine the main conclusions. Nonetheless, at the end of Section 3.5 we discuss how the results might change if this assumption is relaxed.

This study may also provide some insights to the debates on physician dual practice in Canada. In early 2006, the provincial government of Alberta proposed a health reform plan targeting at joint replacement and cataract surgeries.[3] The key of the plan was to allow physician dual practice and allow patients to use private financing to pay for private surgeries. Government officials of Alberta argued that the plan was to offer more options to patients and reduce public waiting times. But opponents criticized that the plan would worsen the already long public waiting lists. Our study

[1] The modeling of multiple classes of customers in a single queue in face of competition between health providers is in contrast to a single class queue in Chen and Wan (2003), Chen and Wan (2005) and Hassin and Haviv (2003).

[2] See Haddrill, Marilyn. "Cataract Surgery Cost." Retrieved on September 11, 2011 from: http://www.allaboutvision.com/conditions/cataract-surgery-cost.htm.

[3] See CBC News. "Alberta's 'third way' could mean health-care showdown with Ottawa". Retrieved on September 11, 2011 from: http://www.cbc.ca/news/canada/story/2006/02/28/thirdway060228.html (February 28, 2006).

shows that the impact of allowing physician dual practice may be two-fold. Allowing physician dual practice may increase the waiting time of patients in the public queue of dual-practice physicians, but some patients in this queue may also benefit from enhanced service qualities. However, these results are conditional on the assumption that there is a fixed supply of physicians whether physician dual practice is allowed or not. Further investigations are needed if this assumption were relaxed.

The remainder of this chapter is organized as follows. Section 3.2 sets up the model. Section 3.3 derives the simultaneous equations that characterize equilibrium arrival rates and proves the existence of equilibrium. Section 3.4 studies the comparative statics of equilibrium arrival rates and equilibrium waiting times. Section 3.5 discusses the impact of physician dual practice on patient's waiting time and welfare. Section 3.6 collects the results for a minor but relevant case when the service quality of public-only physicians is higher than that of dual-practice physicians. Section 3.7 concludes and provides extended discussions.

3.2 Problem Formulation

This section provides modeling details of waiting time and patient choice. As stated in the introduction, we assume that there are two sets of health providers in the market, i.e., a set of n_d dual-practice physicians and another set of n_a public-only physicians. Each set of health providers serve a separate stream of incoming patients. The service times of two types of physicians may not follow the same distribution. The dual-practice physicians have a service quality Q_d and the public-only physicians have a service quality Q_a. Both Q_d and Q_a are expressed in monetary units. Note that although cataract surgery is a standard procedure, the perceived quality of a service could be related to a health provider's reputation, expertise, the amount of personal attention given to a patient, hospital environments, availability of equipments, the supporting staffs, etc (see Conner-Spady et al. (2008)). The concept of "quality" here abstracts all service attributes except for the waiting time, and thus Q_d and Q_a may not be equal. It is beyond the scope of this study to model how the service quality differentiation is developed. Instead, we take the service quality differentiation as given in our modeling. However, we do not predetermine the ordering of Q_d and Q_a one way or the other.

For elective surgeries, patients do not physically stand in a queue to wait for service, but they do appear as a sequence of jobs in physician's working list. Therefore, we would think of patients as joining virtual queues. In the queue of dual-practice physicians, patients who opt for private care need to pay a price p and are given priority over patients who seek public care for free. As explained in the introduction, the assumption of pure priority follows from the observed waiting time difference in the surgeries of dual-practice physicians. Therefore, there are two classes of patients in the queue of dual-practice physicians. There is only one single class of patients in the queue of public-only physicians. Each class of patients form a separate queue: queue 0 for class 0 patients who seek private care, queue 1 for class 1 patients who seek public care of dual-practice physicians, and queue 2 for class 2 patients who seek public care of public-only physicians. Within a class, patients are served under first-in-first-out (FIFO) service discipline.

Both class 0 and class 1 patients receive services of quality Q_d, and class 2 patients receive services of quality Q_a. Let λ_0, λ_1 and λ_2 be the arrival rates of class 0, class 1 and class 2 patients respectively. Generalized from the results of $M/G/n$ queue with non-preemptive priority (see Appendix B.1), the expected waiting times of class 0 and class 1 patients can be written as functions $w_0(\lambda_0, \lambda_1)$ and $w_1(\lambda_0, \lambda_1)$ respectively. Waiting times w_0 and w_1 have the following functional properties: $\frac{\partial w_0}{\partial \lambda_0} > 0$, $\frac{\partial w_0}{\partial \lambda_1} > 0$, $w_0(0,0) = 0$, $\lim\limits_{\lambda_0 \to \frac{n_d}{m_d}} w_0 = +\infty$; $\frac{\partial w_1}{\partial \lambda_0} > 0$, $\frac{\partial w_1}{\partial \lambda_1} > 0$, $w_1(0,0) = 0$, $\lim\limits_{\lambda_0+\lambda_1 \to \frac{n_d}{m_d}} w_1 = +\infty$, where m_d is the mean service time of dual-practice physicians. Furthermore, we have $\frac{\partial w_1}{\partial \lambda_0} > \frac{\partial w_0}{\partial \lambda_0}$ and $\frac{\partial w_1}{\partial \lambda_1} > \frac{\partial w_0}{\partial \lambda_1}$. Generalized from the results of $M/G/n$ queue, the expected waiting time of class 2 patients can be written as function $w_2(\lambda_2)$. w_2 has the following functional properties: $\frac{\partial w_2}{\partial \lambda_2} > 0$, $w_2(0) = 0$ and $\lim\limits_{\lambda_2 \to \frac{n_a}{m_a}} w_2 = +\infty$, where m_a is the mean service time of public-only physicians. It should be noted that the assumption of a single $M/G/n$ queue for either type of physicians is based on the fact that a patient always has the right to be switched to a different surgeon anytime in the course of treatment. However, if the switching cost is high, the public health system would operate like a system of multiple separate $M/G/1$ queues, one queue for each surgeon. As $M/G/1$ queue is a special instance of $M/G/n$ queue, the results obtained in this study can be readily extended to the system of multiple separate $M/G/1$ queues.

We assume that all queues are subject to a common stream of potential patient arrivals with rate Λ (patient population). Patients are heterogenous with respect to their time costs, denoted by h, which has a cumulative probability function $F(x), x \in [0, +\infty)$. All patients act in a fashion so as to maximize their own benefits. Before making decisions to join a queue, patients already have full knowledge of p, Q_d, Q_a, h (their own time costs and the distribution of h) and the expected waiting time of each service alternative. The exact time a patient needs to wait for a service is known only after the booking of operating theater time is confirmed, so no precise waiting time information can be provided to the patient when she is offered different service alternatives. However, the expected waiting time of each service alternative can be reasonably estimated. For instance, the recent waiting time statistics of a surgeon in public/private clinics could be acquired through word of mouth or public waiting time information release. Therefore, it is reasonable to assume that patients make their decisions based on expected waiting time other than actual waiting time.

If a patient decides to join a queue, she then receives one unit of service of quality Q and thus $Q \cdot 1$ is the gross benefit of the service counter. Therefore, Q can also denote the gross service benefit of a queue. After taking into account the price and the cost of waiting, $(Q - p - hw)$ would be the net service benefit of joining a queue. A patient would choose to join a queue that offers the maximum net benefit or choose to receive outside service, which has a reservation net benefit of zero. The aggregation of patient choices would determine the arrival rates λ_0, λ_1 and λ_2, which in turn determine the expected waiting time of each queue. Therefore, both w and λ are endogenous. In an equilibrium, if exists, patients have no incentive to switch to another queue.

3.3 Equilibrium Arrival Rates and the Existence and Uniqueness of Equilibrium

In this section, by assuming that physician dual practice is allowed, we establish the equilibrium arrival rates and the existence of equilibrium when a set of parameters p, Q_d, Q_a, n_d, n_a, h and Λ are given. We use a general framework to derive the necessary conditions for the existence of equilibrium, and then use the necessary conditions to obtain the system of simultaneous equations that characterize the equilibrium arrival rates. The derivation of equilibrium arrival rates is to express λ_0, λ_1 and λ_2 as (implicit) functions of the given set of parameters. Since waiting times are functions of arrival rates, once λ_0, λ_1 and λ_2 are determined, w_0, w_1 and w_2 are determined. Inversely, arrival rates are functions of waiting times, so once w_0, w_1 and w_2 are determined, λ_0, λ_1 and λ_2 are determined as well. The proofs of relevant lemmas, propositions and theorems of this section are in Appendix B.2 and B.3.

3.3.1 Necessary Conditions of Equilibrium

In the analysis of equilibrium, we find that the sorting of patients into different queues depends on the ordering of the rewards of queues. The reward of a queue is defined as the gross service benefit Q minus the price p (if any). For example, if a set of exogenous parameters are such that $Q_d > Q_a$ and $p > Q_d - Q_a$, then the ordering of queues in terms of rewards, is Q_d (queue 1) $> Q_a$ (queue 2) $> Q_d - p$ (queue 0). However, if the parameters are such that $Q_d > Q_a$ and $p < Q_d - Q_a$, then the ordering of queues is Q_d (queue 1) $> Q_d - p$ (queue 0) $> Q_a$ (queue 2). In order to use a unifying framework to derive the necessary conditions of equilibrium, we use R to denote the reward of a queue. We use subscript **A**, **B** and **C** to denote the classes of patients and the corresponding queues with the highest, medium and lowest rewards respectively. We define "mid" to be the operator that takes the medium value of three arguments, then we must have $R_A = \max\{Q_d,\ Q_a\}$, $R_B = \text{mid}\{Q_d,\ Q_d - p,\ Q_a\}$, $R_C = \min\{Q_d - p,\ Q_a\}$ and $R_A \geq R_B \geq R_C$. Let λ_A, λ_B and λ_C be the arrival rates and w_A, w_B and w_C be the expected waiting times of class A, B and C patients respectively, then we have the following proposition:

Proposition 3.1 *Given $p \geq 0$, in an equilibrium (if exists), one of the following three cases must prevail:*

(i) $w_A \geq w_B \geq w_C$;

(ii) $w_B \geq w_A \geq w_C$, *in which case* $Q_a > Q_d$, $w_B = w_1(\lambda_0, \lambda_1)$ *and* $\lambda_B = \lambda_1 = 0$;

(iii) $w_A \geq w_C \geq w_B$, *in which case* $Q_d > Q_a$ *and* $p > Q_d - Q_a$, $w_A = w_1(\lambda_0, \lambda_1)$, $w_B = w_2(\lambda_2)$, $w_C = w_0(\lambda_0, \lambda_1)$, $\lambda_A = \lambda_1 = \Lambda$, $\lambda_B = \lambda_2 = 0$ *and* $\lambda_C = \lambda_0 = 0$.

Case 1 of Proposition 3.1 states that in equilibrium (if exists), a patient class with higher reward must have a longer waiting time than a patient class with lower reward. Manitoba evidence shows that the public patients of dual-practice physicians have the longest waiting time (23 weeks), followed by the public patients of public-only physicians (15 weeks) and the private patients have the shortest waiting time (4 weeks), i.e., we observed $w_1 > w_2 > w_0$. According to Proposition 3.1, we must have $R_A = Q_d > R_B = Q_a > R_C = Q_d - p$. In this case, the service quality of dual-practice physicians must be higher than that of public-only physicians. DeCoster et al. (2000) point out that dual-practice physicians are high volume physicians that specialize in cataract surgeries. Some attributes of these physicians may be perceived by patients as higher service qualities. For instance, it is noted that the greatest single factor in the success rate of cataract extraction procedures is the volume of operations that a surgeon performs.[4] By $Q_a > Q_d - p$, we have $p > Q_d - Q_a$, which implies that the \$1,000 facility fee is greater than the service quality difference between dual-practice physicians and public-only physicians. Only those patients who are highly sensitive to waiting will choose to pay the premium to access the fastest service.

Case 2 of Proposition 3.1 is a special case when no public patients are served by the dual-practice physicians, i.e., $\lambda_1 = 0$. Since there are no patients in the public queue of dual-practice physicians, w_1 should be interpreted as the virtual waiting time: namely, should one class 1 patient joins, his expected waiting time would be w_B. Case 3 of Proposition 3.1 is another special case when all patients are served in the public queue of dual-practice physicians, i.e., $\lambda_1 = \Lambda$. Again, as there are no patients in the private queue of dual-practice physicians, w_0 should be interpreted as the virtual waiting time. Case 2 and Case 3 of Proposition 3.1 are not essential to our research questions as we are interested in studying equilibria in which none of the three queues is empty. Therefore, we limit our analysis to Case 1 of Proposition 3.1 in the rest of this chapter.

3.3.2 Simultaneous Equations of Equilibrium Arrival Rates

In this subsection, we derive simultaneous equations of equilibrium arrival rates and equilibrium waiting times. Appendix B.2 provides more details on the development of these equations and we only summarize the results here.

Given a set of parameters Q_d, Q_a and p and a set of values w'_A, w'_B and w'_C that follow $w'_A \geq w'_B \geq w'_C$, we can show whether or not they constitute an equilibrium.

[4] See Frey, Rebecca. "Extracapsular Cataract Extraction." Gale Encyclopedia of Surgery: A Guide for Patients and Caregivers. 2004. Retrieved on September 11, 2011 from Encyclopedia.com: http://www.encyclopedia.com/doc/1G2-3406200162.html.

By Proposition 3.1, we can determine the patient type that is indifferent between each pair of queues. For instance, the patient type indifferent between queue A and queue B, denoted by h_{AB}, must satisfy $R_A - h_{AB} \cdot w'_A = R_B - h_{AB} \cdot w'_B$ and thus we have $h_{AB} = \frac{R_A - R_B}{w'_A - w'_B}$. Patients with time cost $h \leq h_{AB}$ prefer queue A to queue B; Patients with time cost $h \geq h_{AB}$ prefer queue B to queue A. Similarly we can determine the indifferent patient type between joining a specific queue and choosing an outside service with zero reservation net benefit. By having these indifferent patient types, we are able to sort the heterogenous patients into four segments as follows:

$$\begin{cases} \text{Joining queue A,} & h \in \left[0, \frac{R_A - R_B}{w'_A - w'_B}\right]; \\ \\ \text{Joining queue B,} & h \in \left[\frac{R_A - R_B}{w'_A - w'_B}, \frac{R_B - R_C}{w'_B - w'_C}\right]; \\ \\ \text{Joining queue C,} & h \in \left[\frac{R_B - R_C}{w'_B - w'_C}, \frac{R_C}{w'_C}\right]; \\ \\ \text{Choosing outside service,} & h \in \left[\frac{R_C}{w'_C}, +\infty\right]. \end{cases} \tag{3.1}$$

Since patient type follows cumulative probability distribution function $F(x)$, we can compute the expected arrival rate to each queue. For instance, the arrival rate to queue A is computed as $\lambda'_A = F\left(\frac{R_A - R_B}{w'_A - w'_B}\right)\Lambda$. Similarly we can compute λ'_B and λ'_C. Inputting λ'_A, λ'_B and λ'_C into waiting time functions $w_0(\lambda_0, \lambda_1)$, $w_1(\lambda_0, \lambda_1)$ and $w_2(\lambda_2)$, we obtain w_A, w_B and w_C. If it turns out that $w'_A = w_A$, $w'_B = w_B$ and $w'_C = w_C$, then the set of w'_A, w'_B and w'_C is a candidate of equilibrium waiting times; otherwise, they should not be equilibrium waiting times. Meanwhile, the analysis described above also provides λ'_A, λ'_B and λ'_C as the candidates for equilibrium arrival rates. In summary, the equilibrium arrival rates and equilibrium waiting times must satisfy the following simultaneous equations:

$$\lambda_A = F\left(\frac{R_A - R_B}{w_A - w_B}\right)\Lambda, \tag{3.2}$$

$$\lambda_B = \left[F\left(\frac{R_B - R_C}{w_B - w_C}\right) - F\left(\frac{R_A - R_B}{w_A - w_B}\right)\right]\Lambda, \tag{3.3}$$

$$\lambda_C = \left[F\left(\frac{R_C}{w_C}\right) - F\left(\frac{R_B - R_C}{w_B - w_C}\right)\right]\Lambda. \tag{3.4}$$

where

$$
\begin{aligned}
w_A &= \begin{cases} w_1(\lambda_0,\lambda_1), & if \quad Q_d \geq Q_a; \\ w_2(\lambda_2), & if \quad Q_d < Q_a. \end{cases} \\
w_B &= \begin{cases} w_0(\lambda_0,\lambda_1), & if \quad Q_d \geq Q_a, p \leq Q_d - Q_a; \\ w_2(\lambda_2), & if \quad Q_d \geq Q_a, p > Q_d - Q_a; \\ w_1(\lambda_0,\lambda_1), & if \quad Q_d < Q_a. \end{cases} \\
w_C &= \begin{cases} w_2(\lambda_2), & if \quad Q_d \geq Q_a, p \leq Q_d - Q_a; \\ w_0(\lambda_0,\lambda_1), & if \quad Q_d \geq Q_a, p > Q_d - Q_a \quad \text{or} \quad Q_d < Q_a. \end{cases}
\end{aligned}
$$

Simultaneous equations (3.2) – (3.4) with three unknowns λ_A, λ_B and λ_C serve as the necessary conditions for the existence of equilibrium. Solving (3.2) – (3.4) for λ_A, λ_B and λ_C gives rise to multiple solutions, but only those solutions that satisfy $\lambda_A \geq 0, \lambda_B \geq 0, \lambda_C \geq 0$ are equilibrium arrival rates. Therefore, equations (3.2) – (3.4) with non-negativeness of solutions serve as the sufficient conditions of equilibrium.

According to (3.1), patients choose among service alternatives by weighing their opportunity costs of time against the benefits of service they receive. Patients with lower time costs would choose to receive a higher service benefit by sacrificing their time in waiting. On the other hand, patients who have higher time costs would choose to receive faster service by forgoing service of higher benefit.

3.3.3 Existence and Uniqueness of Equilibrium

This subsection discusses the existence and uniqueness of equilibrium. For the existence of equilibrium, we have the following proposition:

Proposition 3.2 *Given $p > 0$, there always exists an equilibrium.*

Uniqueness of equilibrium is hard to establish in general. For instance, we show an example of multiple equilibria in Appendix B.5. However, for some simple forms of waiting time functions, the uniqueness of equilibrium could be established. Proposition 3.3 presents one of these cases.

Proposition 3.3 *If waiting time functions are of the linear forms $w_0 = \alpha(\lambda_0 + \lambda_1)$, $w_1 = \beta(\lambda_0 + \lambda_1)$ and $w_2 = \alpha\lambda_2$, then the equilibrium is unique.*

By establishing the existence of equilibrium, we extend the study to discuss how the equilibrium arrival rates and equilibrium waiting times respond to the change of price p and the change of service quality difference $Q = |Q_d - Q_a|$ in Section 3.4. In Section 3.5, we study the impact of allowing physician dual practice on patient's waiting time and welfare. As implied in the observations on Manitoba's empirical evidence, the service quality of dual-practice physicians is higher than that of public-only physicians. We therefore focus on the case of $Q_d > Q_a$ in Section 3.4 and Section 3.5 as it is more relevant to the motivation of our research. We present the results for the case of $Q_d < Q_a$ in Section 3.6.

3.4 Comparative Statics of Equilibrium Arrival Rates and Equilibrium Waiting Times with Respect to Price and Service Quality Difference

In this section, we discuss how the equilibrium arrival rates and equilibrium waiting times respond to change of price p and change of service quality difference $Q = Q_d - Q_a$. To derive the comparative statics with respect to price, we take partial derivatives with respect to p at both sides of (3.2), (3.3) and (3.4). We then solve the resulted simultaneous equations for $\frac{\partial \lambda_A}{\partial p}$, $\frac{\partial \lambda_B}{\partial p}$ and $\frac{\partial \lambda_C}{\partial p}$. The expressions of $\frac{\partial \lambda_A}{\partial p}$, $\frac{\partial \lambda_B}{\partial p}$ and $\frac{\partial \lambda_C}{\partial p}$ obtained this way are generally messy and intractable, which do not seem to offer further insights (see Appendix B.4). To simplify the expressions of $\frac{\partial \lambda_A}{\partial p}$, $\frac{\partial \lambda_B}{\partial p}$ and $\frac{\partial \lambda_C}{\partial p}$ and obtain some managerial insights, we limit our analysis to some meaningful specific case. The specific case for our considerations is based on the following assumptions: (I) $\lambda_0 + \lambda_1 + \lambda_2 = \Lambda$; (II) $F(x)$ is uniformly distributed in $[0,1]$; (III) linear forms of waiting time; (IV) the service times of both types of physician follow the same distribution. The justifications for these assumptions are as follows.

Assumption (I) assumes that no patients seek outside service. As explained in Introduction, since the cost of outside service is high, patients who can afford such expensive services only constitute a small fraction of the whole patient population. In other words, ignoring outside service in our model would not undermine the main conclusions. Assumption (II) suits a patient population with no clusterings of patients at certain points of time cost. Assumption (III) is applicable when the traffic intensity of a queue is not close to 1. When the traffic intensity is close to 1, the average waiting time would grow exponentially. We do not observe an exponential growth of waiting time in Manitoba's evidence (see the time series of waiting time in Table 3.1). Therefore, it is reasonably to assume that the traffic intensities of queue 0, queue 1 and queue 2 are not close to 1. Nevertheless, for cases when traffic intensities are high and thus the linear forms of waiting time are deemed as inappropriate, we use numerical analysis to obtain insights. Figure 3.1 to Figure 3.3 are examples of numerical analysis. Findings on these numerical examples supplement the analytical results of Proposition 3.4 and Proposition 3.5.

For assumption (IV), to our best knowledge, there have been no empirical studies investigating the operation times of cataract surgeries. This is probably because cataract surgery has become one of the most common, safest and standard surgeries. The operation of cataract extractions usually lasts less than one hour.[5] Cataract surgeries are usually performed on an outpatient basis and surgeons follow the same standard procedures (Pesudovs and Elliott, 2001). In view of these facts, it is reasonable to assume that the variability of operation times would depend more on other factors, e.g., patient's ocular co-morbidities, than surgeon practice types.

3.4.1 Marginal Equilibrium Arrival Rates of Price and Marginal Equilibrium Waiting Times of Price

Given assumptions (I), (II), (III) and (IV), we summarize the analytical results of marginal equilibrium arrival rates of price and marginal equilibrium waiting times of price as Proposition 3.4 below. The details of analysis are in Appendix B.4.

Proposition 3.4 *Under assumptions of (I), (II), (III), (IV) and $Q_d > Q_a$, it follows that:*
$\frac{\partial \lambda_0}{\partial p} < 0$, $\frac{\partial \lambda_1}{\partial p} > 0$, $\frac{\partial \lambda_2}{\partial p} < 0$, $\frac{\partial w_0}{\partial p} < 0$, $\frac{\partial w_1}{\partial p} < 0$, $\frac{\partial w_2}{\partial p} < 0$, *if* $p > Q_d - Q_a$.

We cannot determine the signs of comparative statics analytically for the case of $p < Q_d - Q_a$. Instead, we conduct extensive numerical analysis for that case. Our numerical analysis uses assumption (I), (II) and (IV). Instead of linear forms of waiting time, we use waiting times of $M/G/n$ queue with non-preemptive priority. We fix the patient population to be 1, i.e., $\Lambda = 1$, and limit the number of dual-practice physicians to be less than or equal to 3, i.e., $n_d \leq 3$. This is because when

[5] See National Eye Institute. "Facts About Cataract." Retrieved on September 11, 2011 from: http://www.nei.nih.gov/health/cataract/cataract_facts.asp

$n_d \geq 3$, the computation of the numerical examples becomes extremely slow and would not finish within the given time limit (60 minutes). The number of public-only physicians is fixed to be 1 to utilize the explicit waiting time function of $M/G/1$ queue. The values of Q_d and Q_a are chosen in such a way that the value of $(Q_d - Q_a)$ varies from 0 to 100. For each pair of Q_d and Q_a, the first and second moments of the service time, i.e., m and m^2, would have to fall within a certain region to ensure that none of the queues has zero arrival rate. All of the numerical examples show that λ_0, λ_1, w_0 and w_1 are monotone with respect to price, which is consistent with Proposition 3.4. Arrival rate λ_2 is not monotone with respect to p. Figure 3.1 shows one of these numerical examples.

Both Proposition 3.4 and Figure 3.1 suggest that a price increase deters patients from seeking private service and thus increases the demand for dual-practice physician's public services. Although the arrival rate of class 1 patients increases in p, the expected waiting time of class 1 patients decreases thanks to the decrease of the arrival rate of class 0 patients.

The interesting finding is how λ_2 respond to p. Proposition 3.4 and Figure 3.1 suggest that different patterns could exist. In the case of linear waiting time, Proposition 3.4 suggests that the arrival rate to queue 2 is monotonically decreasing in price p in the region of $p < (Q_d - Q_a)$. In the case of $M/G/1$ queue, Figure 3.1 shows that when the price of private service is lower than the service quality difference, i.e., $p < (Q_d - Q_a)$, the line of λ_2 is flat, so the price change only induces a switching of patients between queue 0 and queue 1. When the price of private service is higher than the service quality difference, i.e., $p > (Q_d - Q_a)$, Figure 3.1 shows that the line of λ_2 is quasi-concave in p. In this case, patients with low/medium/high time costs would join queue 1/2/0 respectively. An increase of the already high price will make queue 0 less attractive to patients with high time costs, so some of these patients would switch from queue 0 to queue 2. Meanwhile, there are also patients switching from queue 2 to queue 1 because the expected waiting time of class 1 patients decreases. Therefore, the change of the arrival rate of class 2 patients may not be monotone.

3.4.2 Marginal Equilibrium Arrival Rates of Service Quality Difference and Marginal Equilibrium Waiting Times of Service Quality Difference

This subsection also utilizes assumptions (I), (II), (III) and (IV). The analytical results of marginal equilibrium arrival rates of service quality difference and marginal equilibrium waiting times of service quality difference are summarized as Proposition 3.5. The change of service quality difference $Q = Q_d - Q_a$ can be due to either the change of Q_d, or the change of Q_a, or both. In the following discussions, we assume that the increase of Q is due to the increase of Q_d. The details of the analysis are in Appendix B.4.

Proposition 3.5 *Under assumptions (I), (II), (III), (IV) and $Q_a < Q_d$, it follows that:*

(i) $\frac{\partial \lambda_2}{\partial Q} < 0,\ \frac{\partial w_0}{\partial Q} > 0,\ \frac{\partial w_1}{\partial Q} > 0,\ \frac{\partial w_2}{\partial Q} < 0,$ *if* $p > Q_d - Q_a$;

(ii) $\frac{\partial \lambda_0}{\partial Q} > 0,\ \frac{\partial \lambda_1}{\partial Q} < 0,\ \frac{\partial \lambda_2}{\partial Q} < 0,\ \frac{\partial w_0}{\partial Q} > 0,\ \frac{\partial w_1}{\partial Q} > 0,\ \frac{\partial w_2}{\partial Q} < 0,$ *if* $p < Q_d - Q_a$.

We cannot determine the signs of $\frac{\partial \lambda_0}{\partial Q}$ and $\frac{\partial \lambda_1}{\partial Q}$ analytically under assumptions (I), (II), (III) and (IV) when $p > Q_d - Q_a$. Instead, we conduct extensive numerical analysis following the same choices of parameters in Section 3.4.1. Figure 3.2 and 3.3 are two numerical examples.

In the case of linear waiting times, Proposition 3.5 suggests that an increase of service quality of dual-practice physicians makes patients switch from the queue of public-only physicians to the queue of dual-practice physicians. Therefore, the arrival rate of class 2 patients decreases. When the price is less than the service quality difference, i.e., Case 2 of Proposition 3.5, an increase of service quality of dual-practice physicians makes the price p become less concerned to class 1 and class 2 patients. Therefore, class 1 and class 2 patients are more willing to pay to enjoy a service of expediency and higher quality. An increase of the already high Q_d will make patients switch from both queue 1 and queue 2 to queue 0.

The monotonicity stated in Proposition 3.5 holds in the numerical example of Figure 3.2, but fail to hold in the numerical example of Figure 3.3. As noted above, the linear approximation of waiting time is more applicable when the traffic intensity is not close to 1, i.e., the system is not congested. The only difference between Figure

3.2 and Figure 3.3 is the mean service time: $m = 2.22$ for Figure 3.2 and $m = 2.5$ for Figure 3.3. With everything else being equal, a queue with higher mean service time would have higher traffic intensity and thus is more congested. In the more congested example of Figure 3.3, when the service quality difference is lower than the price, the change of λ_0 with respect to Q is no longer monotonic. When the service qualities are close, i.e., Q_d is close to Q_a, patients do not see much value in paying a high price p to access private care. Therefore, patients are more willing to enjoy a higher service quality Q_d for free. Also, patients know that since the system is congested, the waiting time of class 1 patients would drop greatly if the number of class 0 patients falls. By switching from queue 0 to queue 1, patients in queue 1 not only enjoy a higher service quality for free, but also enjoy a shorter waiting time because the traffic intensity of queue 0 is reduced. In fact, the waiting time of class 1 patients does not change much in the region $Q_d \in [2, 2.4]$. Therefore, an increase of service quality of dual-practice physicians not only makes patients switch from queue 2 to queue 1, but also makes patients switch from queue 0 to queue 1.

3.5 Physician Dual Practice on Patient's Waiting Time and Welfare

In this section, we discuss the impact of allowing physician dual practice on patient's waiting time and welfare. We first set up the base case in which physician dual practice is not allowed and patients are unable to distinguish between physicians of different service qualities. We obtain the waiting time, and then calculate the net service benefit of each patient type in the base case. Next, we compare the waiting times and welfare of class 0, 1 and 2 patients with the waiting time and welfare of the base case. In this section, we assume no outside service whether physician dual practice is allowed or not, and we only discuss the case of $Q_d > Q_a$. We will present the results for the case of $Q_d < Q_a$ in Section 3.6.

Physician dual practice could affect patient's welfare through two channels: one is through waiting time differentiation, and the other is through service quality improvement. The existing literature has shown that allowing physician dual practice could improve the service quality of dual-practice physicians in the public health system (Eggleston and Bir, 2006). González (2004) argues that allowing physician dual practice improves the service quality as dual-practice physicians conduct more accurate diagnosis for patients. This is because dual-practice physicians need to establish reputations among patients. Section 3.3 has shown that the waiting time difference of public surgeries existing between two types of physician can be explained by service quality differentiation. This finding motivates us to take both the effect of waiting time differentiation and the effect of service quality enhancement into account, and investigate how these two effects work together to affect patient's welfare. It is beyond the scope of this study to model how the service quality enhancement is realized. Instead, we take the service quality enhancement effect as given in our model.

3.5.1 The Base Case

In this subsection, we establish the waiting time and patient's welfare of the base case – a setting in which physician dual practice is not allowed. In such a setting, patients are unable to distinguish the service quality difference, if any, among physicians because there are no price discrimination or the offering of private service. Meanwhile, physicians cannot price discriminate patients either. A patient arriving at this system joins the end of a common queue and upon reaching the head of the queue, she is assigned to the next available physician. This system can be considered as an $M/G/(n_d+n_a)$ queue under FIFO service principle. Here n_d is the number of physicians who would become dual-practice physicians if physician dual practice is allowed; n_a is the number of physicians who only practice in the public system whether physician dual practice is allowed or not. Since we assume that there is no outside service, all of Λ patients are served in the $M/G/(n_d+n_a)$ queue. The expected waiting time of patients in the base case can be written as $\widetilde{w}(\Lambda, n_d+n_a)$. We assume that the n_d physicians have a service quality of $\widetilde{Q}_d$ and the n_a physicians have a service quality of Q_a in the base case. To model the service quality enhancement effect induced by allowing physician dual practice, we assume $\widetilde{Q}_d \leq Q_d$.

We assume that the total number of physicians is fixed whether physician dual practice is allowed or not. This is a valid assumption in the short term as new physicians would face entry barrier. In Manitoba's case, none of the dual-practice physicians left the public system after extra billing was banned in January 1999. Bir and Eggleston (2003) argue that allowing physician dual practice could help attract higher skill physicians. However, they fail to formalize a concrete model to justify this argument. We therefore treat the number of physicians as fixed in this section for the sake of simplicity. At the end of this section, we discuss the possible impacts of relaxing this assumption.

As patients are randomly assigned to physicians, a patient is expected to receive a service of quality $\widetilde{Q}_d$ with probability $n_d/(n_d+n_a)$ or a service of quality Q_a with probability $n_a/(n_d+n_a)$. The expected net benefit for a patient with time cost h is

$$\frac{n_d\widetilde{Q}_d+n_aQ_a}{n_d+n_a}-h\cdot\widetilde{w}(\Lambda,n_d+n_a)$$

and the total patient welfare of the base case is

$$\Lambda\left(\frac{n_d\widetilde{Q}_d+n_aQ_a}{n_d+n_a}-\widetilde{w}(\Lambda,n_d+n_a)\int_0^{+\infty}h\,dF(h)\right)$$

which would be referred as the benchmark patient welfare in the following discussions.

It should be noted that due to the pooling effect, when a common $M/G/(n_d+n_a)$ queue is split into a $M/G/n_d$ queue and another $M/G/n_a$ queue, some efficiency is lost. This means at least one of the separated queues would have a longer waiting time than the original common queue, regardless how the total arrival rate is split into the two separated queues.

3.5.2 Impact of Physician Dual Practice on Waiting Time

In this subsection, we compare the waiting time of the base case with the waiting times of class 0, class 1 and class 2 patients. By Proposition 3.1, we know that in the equilibrium of case $Q_a<Q_d$, class 1 patients would be the group of patients having lower time costs than class 0 or class 2 patients. One key issue in the health reform debates in Canada is whether allowing privately funded health care would result in a shorter public waiting time. In particular, the policy maker is concerned about whether allowing privately funded health care would benefit some people at the cost of others. Proposition 3.6 below shows that given a fixed supply of physicians, allowing physician dual practice results in a longer waiting time for patients with lower time costs, i.e., class 1 patients.

Proposition 3.6 *Given $p>0$, $Q_a<Q_d$ and a fixed supply of physicians, we always*

have $\widetilde{w} \leq w_1$, i.e. the waiting time of patients with lower time costs becomes longer when physician dual practice is allowed.

There are no similar conclusions for class 0 and class 2 patients. As shown in the right hand sides of Figure 3.1 to Figure 3.3, when $Q_a < Q_d$, w_1 is always longer than the waiting time of the base case (the black bar). However, w_0 and w_2 could be longer or shorter than the waiting time of the base case.

It should be noted that patients with lower time costs are not necessarily patients with lower incomes. As shown in DeCoster et al. (2000), a substantial proportion (38%) of total private surgeries were performed on patients from the two lowest-income neighborhoods. Clearly these patients have higher time costs. Therefore, patient's opportunity cost of time could be determined by factors other than personal income. For instance, in the case of cataract surgery, a patient's cost of time could be mainly determined by the importance of vision to her life and work. Allowing privately funded health service offers more options not only to the rich but also to the poor.

Figure 3.1 to Figure 3.3 show that the proportion of class 1 patients could range from 0% to a large percentage depending on the price and the service qualities. Table 3.2 below shows that nearly 50% of Manitoba cataract patients were class 1 patients during the time period when physician dual practice was allowed.

Figure 3.1: Numerical example with $Q_d = 2.4,\ Q_a = 2,\ \Lambda = 1,\ m = 2.5,\ m^2 = 1,\ n_d = 2,\ n_a = 1$. (Note: the expected waiting time of class 1 patients is plotted against the secondary Y-axis.)

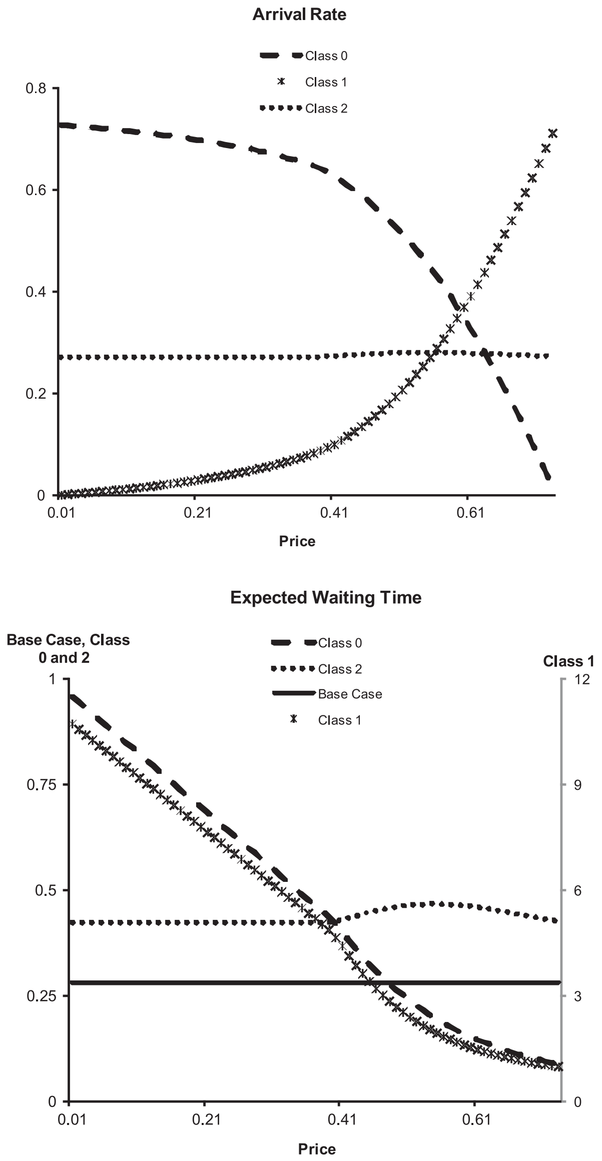


Figure 3.2: Numerical example with $p = 0.4,\ Q_a = 2,\ \Lambda = 1,\ m = 2.22,\ m^2 = 1,\ n_d = 2,\ n_a = 1$. (Note: the expected waiting time of class 1 patients is plotted against the secondary Y-axis.)

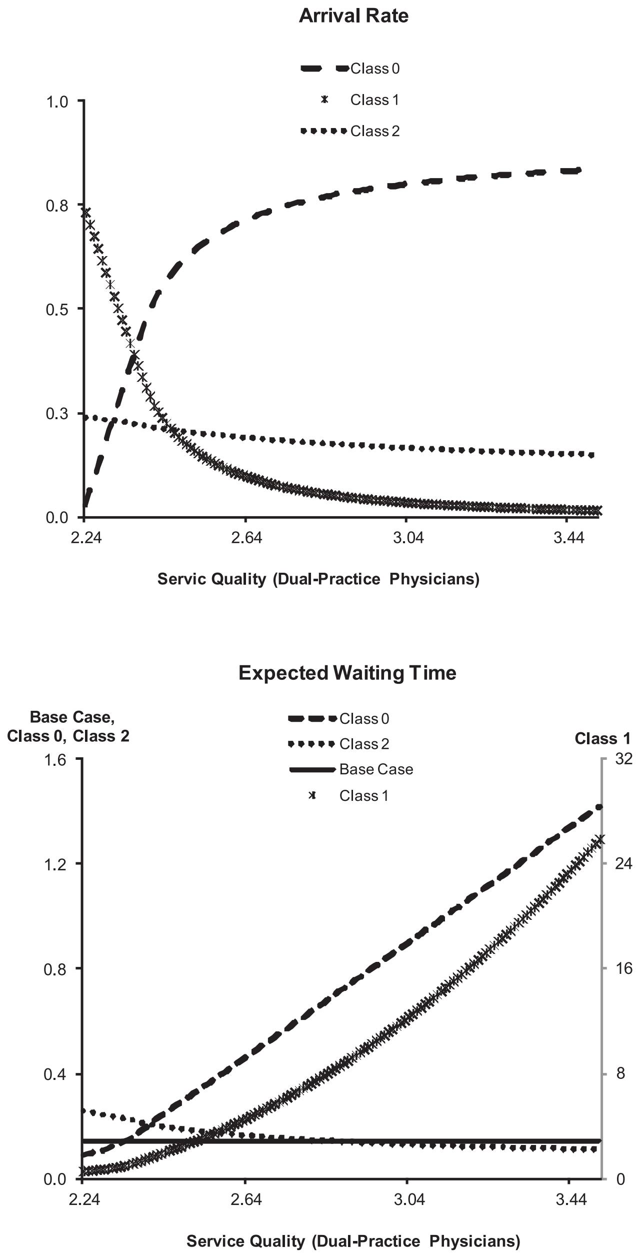


Figure 3.3: Numerical example with $p = 0.4,\ Q_a = 2,\ \Lambda = 1,\ m = 2.5,\ m^2 = 1,\ n_d = 2,\ n_a = 1$. (Note: the expected waiting time of class 1 patients is plotted against the secondary Y-axis.)

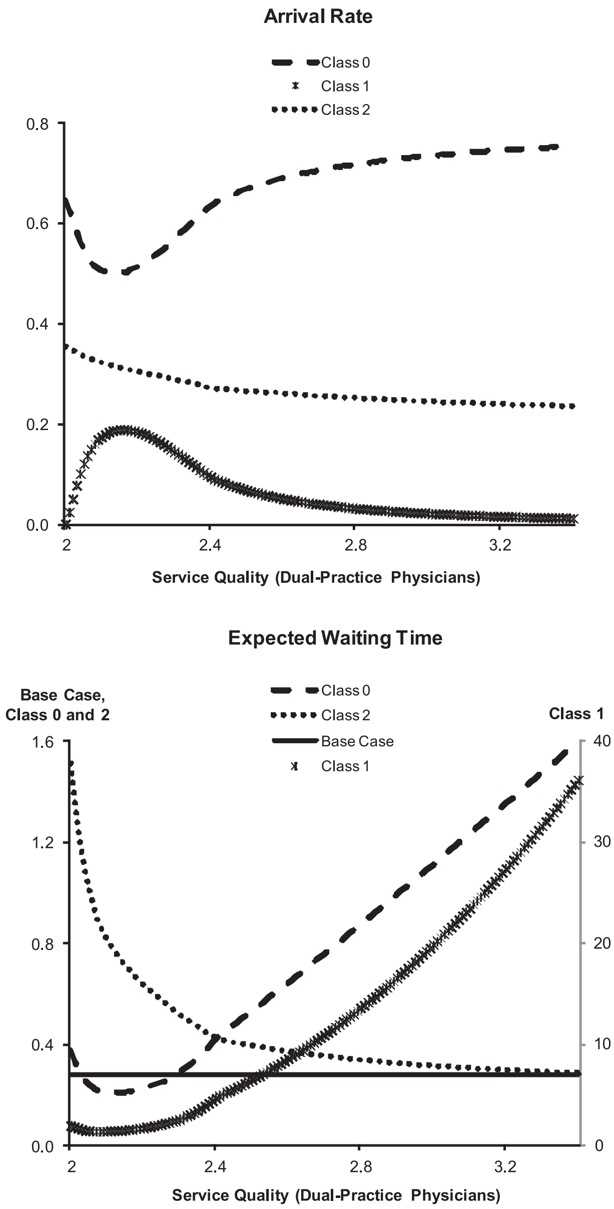


Table 3.2: Number of patients by surgeon practice type: 1992-1999[a]

	Year	92/93	93/94	94/95	95/96	96/97	97/98	98/99
Public-only surgeon	Public hospital	1,207	1,190	1,365	1,578	1,471	1,133	1,154
		(36.0%)	(35.6%)	(40.6%)	(39.4%)	(38.1%)	(25.8%)	(23.4%)
Dual-practice surgeon	Public hospital	1,671	1,705	1,689	2,043	1,722	2,353	2,424
		(49.9%)	(51.0%)	(50.2%)	(50.9%)	(44.6%)	(53.6%)	(49.2%)
	Private clinic	471	448	312	389	672	903	1,351
		(14.1%)	(13.4%)	(9.3%)	(9.7%)	(17.4%)	(20.6%)	(27.4%)
		3,349	3,343	3,366	4,010	3,865	4,389	4,929

[a] Source: DeCoster et al. (2000).

3.5.3 Impact of Physician Dual Practice on Patient Welfare

In the preceding subsection, we have shown that class 1 patients have a longer waiting time than the waiting time of the base case. However, class 1 patients may benefit from the enhanced service quality induced by allowing physician dual practice. Therefore, the net effect of physician dual practice on class 1 patients is not straightforward. Suppose that a patient with time cost h joins queue 1, so she receives a net benefit of $Q_d - hw_1$. This net benefit $Q_d - hw_1$ is no worse than the net benefit she receives in the base case if and only if:

$$\begin{aligned} & Q_d - hw_1 \geq \frac{n_d\widetilde{Q}_d + n_a Q_a}{n_d + n_a} - h\widetilde{w} \\ \Rightarrow \qquad & Q_d - \frac{n_d\widetilde{Q}_d + n_a Q_a}{n_d + n_a} \geq h(w_1 - \widetilde{w}) \end{aligned} \tag{3.5}$$

According to Proposition 3.6, we know that $w_1 > \widetilde{w}$. Dividing both sides of (3.5) by $(w_1 - \widetilde{w})$ gives rise to

$$h \leq \left(Q_d - \frac{n_d\widetilde{Q}_d + n_a Q_a}{n_d + n_a} \right) / \, (w_1 - \widetilde{w}) \tag{3.6}$$

Since $\widetilde{Q}_d < Q_d$ and $Q_a < Q_d$, we have $Q_d - \frac{n_d\widetilde{Q}_d + n_a Q_a}{n_d + n_a} > 0$. The right hand side of (3.6) is always positive, i.e., there always exist class 1 patients who are better off by allowing physician dual practice. Furthermore, we let

$$h_d = \min\{h_1, h_2\}$$

where h_1 denotes the right hand side of (3.6), while h_2 denotes the indifferent type between queue 1 and queue 2 (in the case of $Q_d - Q_a \leq p$) or the indifferent type between queue 1 and queue 0 (in the case of $Q_d - Q_a > p$). Therefore, h_d is a threshold of time cost. Patients with time costs lower than this threshold would benefit from allowing physician dual practice. As both h_1 and h_2 are functions of p and Q_d, h_d is a function of p and Q_d as well, i.e., $h_d(p, Q_d)$. For the marginal h_d of price p, we have the following proposition:

Proposition 3.7 *Given $Q_a < Q_d$, we have $\frac{\partial h_d}{\partial p} > 0$, i.e., the higher the price, the more class 1 patients who would benefit from allowing physician dual practice.*

Proposition 3.7 holds because as price increases, the number of class 1 patients increases and the waiting time of class 1 patients decreases. Therefore, class 1 patients would enjoy a shorter waiting time and a service of higher quality (compared to the base case). As the price increases, more class 1 patients would receive a higher net benefit than they do in the base case. We cannot make a similar conclusion for the marginal h_d of service quality Q_d.

Next, we do the same analysis for class 0 and class 2 patients. A patient with time cost h in queue 0 would benefit from allowing physician dual practice if and only if:

$$Q_d - p - \frac{n_d\widetilde{Q}_d + n_a Q_a}{n_d + n_a} \geq h(w_0 - \widetilde{w}) \tag{3.7}$$

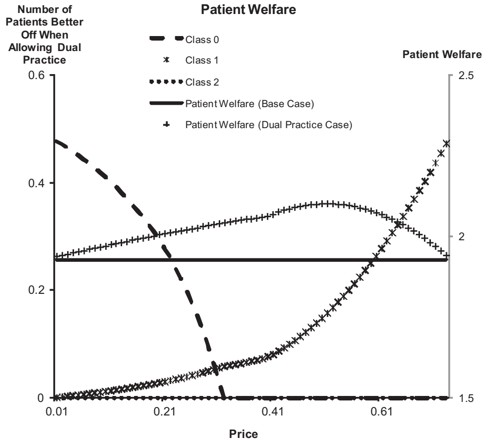
Both the sign of $Q_d - p - \frac{n_d\widetilde{Q}_d + n_a Q_a}{n_d + n_a}$ and the sign of $(w_0 - \widetilde{w})$ could be either positive or negative. The discussion would lead to many scenarios. Likewise, a patient with time cost h in queue 2 would benefit from allowing physician dual practice if and only if:

$$Q_a - \frac{n_d\widetilde{Q}_d + n_a Q_a}{n_d + n_a} \geq h(w_2 - \widetilde{w}) \tag{3.8}$$

Again, both the sign of $Q_a - \frac{n_d\widetilde{Q}_d + n_a Q_a}{n_d + n_a}$ and the sign of $(w_2 - \widetilde{w})$ could be either positive or negative, and the discussion would lead to many scenarios too. In fact, due to the loss of efficiency in queue splitting, we can construct an example in which none of the class 0 and class 2 patients is better off from allowing physician dual practice.

Figure 3.4 is a numerical example using the same parameters as Figure 3.1. Figure 3.4 shows how the number of class 0/1/2 patients who would benefit from allowing physician dual practice changes with respect to price. The number of queue 0/1/2 patients who would benefit from allowing physician dual practice is labeled as line “queue 0/1/2”. Figure 3.4 also shows the total patient welfare of the base case and that of the dual practice case. As shown in Figure 3.4, no class 2 patients would benefit from allowing physician dual practice due to the loss of efficiency in queue splitting. The single public-only physician serves less patients in the case when physician dual practice is allowed, but the waiting time of class 2 patients is instead longer than the waiting time of the base case. The number of class 1 patients who would benefit from allowing physician dual practice increases in price. The number of class 0 patients who would benefit from allowing physician dual practice decreases in price. This is because as the price increases, less patients would stay in queue 0 and those patients who remain staying in queue 0 need to pay a higher price. The total patient welfare is quasi-concave in price. The total patient welfare is at the lowest when the price is either very low or very high. It should be noted that the comparison of total patient welfare between the base case and the case where physician dual practice is allowed is conditional on the choice of $\widetilde{Q}_d$. Therefore, we should refrain from making conclusions on which scenario yields higher total welfare.

Figure 3.4: Patient welfare with $Q_d = 2.4,\ Q_a = 2,\ \widetilde{Q}_d = 2.1,\ \Lambda = 1,\ m = 2.5,\ m^2 = 1,\ n_d = 2,\ n_a = 1.$



To study the change of patient welfare with respect to service quality, we can fix Q_a and Q_d but allow $\widetilde{Q}_d$ to vary. As $\widetilde{Q}_d$ increases, the service quality enhancement, i.e., $(Q_d - \widetilde{Q}_d)$, would decrease. For the number of class 0/1/2 patients who would benefit from allowing physician dual practice with respect to $\widetilde{Q}_d$, we have the following proposition:

Proposition 3.8 *Given $Q_a < Q_d$, the higher the service quality $\widetilde{Q}_d$ in the base case, the less the patients across all classes would benefit from allowing physician dual practice.*

The intuition behind Proposition 3.8 is straightforward: the extent of patient welfare improvement correlates positively with the extent of service quality improvement induced by physician dual practice.

3.5.4 Discussion

In the above subsections, we have relied on two important assumptions to simplify the analysis: one assumption is no outside service and the other is a fixed number of physicians whether physician dual practice is allowed or not. In this subsection, we discuss the possible impacts if these two assumptions are relaxed.

In the setting of this study, an outside service means a health service that is offered in another jurisdiction. For instance, an outside service for cataract patients in Manitoba could be a cataract surgery that is offered in a neighborhooding province like Ontario, or a cataract surgery that is offered across the border. To model an outside service option, the most simplistic way is to assume that the outside service offers a reservation net benefit v to patients regardless of their types. A patient would choose the outside service if she finds that the maximum net benefit from being served by either the dual-practice physicians or the public-only physicians is lower than v. Our discussion below relies on this modeling approach.

In the base case, the introduction of an outside service provides patients an alternative if they feel that the public waiting time is long. The outside service functions as a valve to mitigate the congestion in the public system, and thus the introduction of an outside service would inevitably shorten the public waiting time and improve patient's welfare. In the case when physician dual practice is allowed, the introduction of an outside service could lead to mixed outcomes on patient's waiting time. Appendix B.5 has shown an example of multiple equilibria in the case when physician dual practice is allowed. This example shows that some patients would switch from the existing system to the outside service if there were one. However, whether or not the introduction of outside service would improve patient's waiting time really depends on which equilibrium the system would be driven into. In short, the introduction of an outside service always favors the base case.

Meanwhile, we may also be interested in whether the main results of this section would hold if there were an outside service whether physician dual practice is allowed or not. With an outside service in the market, class 1 patients would have a longer waiting time in the case when physician dual practice is allowed.[6] This means Proposition 3.6 would still hold. Similarly, there always exist class 1 patients who would benefit from allowing physician dual practice. However, Proposition 3.7 and

[6] One can prove it by using the same logic as the proof of Proposition 3.6

Proposition 3.8 may or may not hold.

As stated above, Bir and Eggleston (2003) argue that allowing physician dual practice could help attract higher skilled physicians. Indeed, allowing physician dual practice introduces additional possible venues for physicians to supplement their public incomes, so some physicians from outside might be attracted to the system and thus the total number of physicians increases. Intuitively, one might expect that once the assumption of fixed number of physicians is relaxed, the the waiting time of at least one of the patient classes would decrease. This conjecture seems reasonable because there is more service supply in the market. However, there might exist some counterexamples to this seemingly logic argument. For instance, Appendix B.5 shows that the waiting time of each patient class in Table B.3 is longer than the waiting time of the same patient class in Table B.2, even if there are same number of dual-practice physicians and same number of public-only physicians in both cases. This example also shows that the patient welfare in Table B.3 is no better than that in Table B.2. That is, the equilibrium of Table B.3 is dominated by the equilibrium of Table B.2 in terms of patient welfare. Think of a case in which the addition of a negligible number of physicians would trigger the system to evolve from the equilibrium of Table B.2 to the equilibrium of Table B.3. In that case, the addition of physician supply does not improve patient's welfare. Therefore, it is not straightforward how relaxing the assumption of fixed number of physicians would affect the main results of this section. We leave it as a direction for future research.

3.6 Results for the Case of $Q_d < Q_a$

We assume $Q_d < Q_a$ throughout this section, i.e., the service quality of dual-practice physicians is lower than that of public-only physicians. Although this case is not the main focus of this chapter, we summarize the results here to supplement the findings presented in the previous sections. We first present the comparative statics of equilibrium arrival rates and equilibrium waiting times with respect to price and service quality difference. We then present the results as how allowing physician dual practice would affect patient's waiting time and welfare.

Given assumption (I), (II), (III) and (IV), the analytical results of marginal equilibrium arrival rates and marginal equilibrium waiting times are summarized as Proposition 3.9 below. The details of the analysis are in Appendix B.4. We assume that the change of service quality difference $Q = Q_a - Q_d$ is due to the increase of Q_a.

Proposition 3.9 *Under assumptions (I), (II), (III), (IV)and $Q_d < Q_a$, it follows that:*

(i) $\frac{\partial \lambda_0}{\partial p} < 0$, $\frac{\partial \lambda_1}{\partial p} > 0$, $\frac{\partial \lambda_2}{\partial p} = 0$, $\frac{\partial w_0}{\partial p} = 0$, $\frac{\partial w_1}{\partial p} = 0$, $\frac{\partial w_2}{\partial p} = 0$;

(ii) $\frac{\partial \lambda_0}{\partial Q} < 0$, $\frac{\partial \lambda_1}{\partial Q} < 0$, $\frac{\partial \lambda_2}{\partial Q} > 0$, $\frac{\partial w_0}{\partial Q} < 0$, $\frac{\partial w_1}{\partial Q} < 0$, $\frac{\partial w_2}{\partial Q} > 0$.

Case 1 of Proposition 3.9 shows that a price increase deters patients from seeking private care and thus increases the demand for dual-practice physicians' public service. However, due to the linear forms of waiting time, the arrival rate to queue 2, as well as the waiting times of all three patient classes, does not change in price. Numerical analysis[7] instead shows that the arrival rate to queue 2, as well as the waiting times of all three patient classes, is decreasing in p. Case 2 of Proposition 3.9 shows that an increase of the service quality of public-only physicians attracts patients from both queue 0 and queue 1, and thus both the arrival rate and waiting time of class 2 patients increase.

With regard to the impact of allowing physician dual practice on patient's waiting time and welfare, we summarize the results as Proposition 3.10:

Proposition 3.10 *Given $Q_a > Q_d$ and $p > 0$, we have*

(i) $\widetilde{w} < w_2$, *i.e., the waiting time of patients with lower time costs becomes longer when physician dual practice is allowed;*

(ii) *There always exist class 2 patients who would benefit from allowing physician dual practice;*

(iii) *The higher the service quality $\widetilde{Q}_d$ in the base case, the less the class 0/1/2 patients who would benefit from allowing physician dual practice.*

Both the conclusion and proof of Proposition 3.10 follow those of Proposition 3.6 to Proposition 3.8, so details are skipped here for the sake of conciseness.

[7] The numerical analysis here follows the same choices of parameters as in Section 3.4.1 except that the values of Q_d and Q_a are chosen in such a way that the value of $(Q_a - Q_d)$ varies from 0 to 100.

3.7 Concluding Remarks

Physician dual practice is a common phenomenon in many OECD and developing countries (García-Pardo and González, 2007; Jan et al., 2005), but the existing literature on physician dual practice is relatively limited and recent (Eggleston and Bir, 2006; González, 2005). Motivated by the two types of waiting time differences existing in Manitoba's cataract surgeries, we study physician dual practice from a different but related perspective. In this chapter, we showed that the waiting time difference existing between dual-practice physicians and public-only physicians may be explained by the service quality differentiation between the two groups of physicians. Patients of physicians with higher service qualities have longer waiting times than patients of physicians with lower service qualities. The impact of allowing physician dual practice on patient's waiting time and welfare is mixed. We showed that allowing physician dual practice would increase the waiting time for patients with lower time costs, but some of these patients may also benefit from an enhanced service quality induced by allowing physician dual practice. The impact of allowing physician dual practice on patients with higher time costs is mixed.

Chapter 4

Tax or Subsidy on Private Care and Income Redistribution in A Two-Tier Health System

4.1 Introduction

Public provision of private goods, e.g., health care and eduction, has long been a critical subject in public economics. Governmental inventions into markets of such goods and service are deemed as needed due to a number of reasons. First, free market by its own can be inefficient under certain circumstances: monopoly power of suppliers, negative externalities resulted from adverse selection and moral hazard, and asymmetric information. Second, because of the market failure in the private insurance market, a social insurance or a tax-financed health system may be more efficient than a competitive private insurance system. Last, public provision of private goods serves redistributive purpose, i.e., redistribution of wealth and income from persons with higher earning abilities to persons with lower earning abilities. For a comprehensive coverage of the arguments for public provision of health care, please see Section 8 of McPake et al. (2002), Section 3 of Wonderling et al. (2005) and Section 5 of Zweifel et al. (2009). In this study, we focus on the redistributive purpose of the public provision of health care.

One may argue that the redistributive purpose can be accomplished by income taxation alone. However, one of the lessons learned from the optimal income tax literature is that there are limits to the amount of redistribution that can be achieved by progressive taxation (e.g., Mirrlees (1971); Roberts (1971)). Personal income is not a perfect signal of earning ability as income could depend on labor supply and human capital investments (Besley and Coate, 1991). Another constraint in designing an efficient tax-based redistribution policy is that governments are not as well-informed about the relevant utility-determining characteristics of taxpayers as the taxpayers themselves are (Boadway et al., 1998). Public provision of private goods, on the other hand, could be used as a policy instrument to reduce these inefficiencies (e.g., Bucovetsky (1984); Blomquist and Christiansen (1995); Boadway and Marchand (1995)). For instance, Cremer and Gahvari (1997) show that as a redistributive mechanism, public provision of private goods can enhance overall welfare above the maximum that can be achieved when the income tax policy is designed optimally on the basis of the information available to the government. Besley and Coate (1991) show that universal provision of private goods of certain quality levels can serve the redistributive purpose even if the provision is financed by a head tax. In this chapter, we show that public provision of health care can further improve income redistribution if an optimally designed subsidy or tax (i.e., "negative subsidy") were levied on private care when the public provision is financed by a head tax. We then characterize the conditions under which providing subsidy to private care improves income redistribution.

The key assumption in Besley and Coate (1991) to drive their results is that the quality levels of public and private provisions must be sufficiently differentiated to make consumer's self-selection to be an informal means for efficient sorting of consumers between sectors. In this chapter, the quality differentiation comes from patient's waiting. Bucovetsky (1984) and Hoel and Sáther (2003) show that waiting can be an efficient method for income redistribution. The argument rests on the difference in the valuation of time among different patients – whereas the private providers use this difference to price discriminate, the welfare-maximizing government uses it as a redistributive tool (Bucovetsky, 1984). Chapter 2 has shown that providing subsidy to private care contributes to public waiting time reduction. An important and related question to this finding is: Whether or not the policy of providing subsidy would improve patient's income redistribution. This chapter shows that the answer to this question is conditional. When the utilization[1] of the public health system is

[1] The utilization is defined as the ratio of total patient arrival rate to the service rate of the public

high, the public waiting time is expected to be long and patients with high time costs would seek private care. A tax levied on the consumption of private care would generate a revenue transferred from the private sector to the public health system. The transferred revenue would then reduce the cost of public care that everyone pays for (i.e., the head tax). On the contrary, when the utilization is low, the public waiting time is expected to be short, so patients with high time costs are likely to stay in the public health system. The production cost of public care would then be high and so would be the head tax. To induce the patients with high time costs to the private sector, a subsidy to private care should be provided. This result agrees the conclusion of Eggleston and Bir (2006) that by inducing higher income consumers to the private sector, public care becomes more effectively targeted on poor consumers.

Similar to Hoel and Sáther (2003), the consideration of income redistribution is modeled as the weights that the health planner assigns to the welfare of different patient types. Using welfare weights to model government's redistributive objective is an approach used in optimal income taxation literature (e.g., Boadway et al. (1998)). Patients with high earning abilities have high opportunity costs of time. These patients are more sensitive to the waiting in the public health system. Therefore, the health planner is more concerned about the welfare of patients with low time costs and thus assigns proportionally higher weights to them. Our study shows that the more the health planner assigns weights to patients with low time costs, the more likely the health planner implements tax on private care.

The model we use in this chapter is a straightforward extension of the one in Hoel and Sáther (2003). Nevertheless, to consider the effect of tax or subsidy on patient's income redistribution, patient's waiting can no longer be treated as an independent decision variable as in Hoel and Sáther (2003). The health planner's tax or subsidy decision would influence patient's choice, which determines the effective arrival rate to the public health system and in turn determines the public waiting time. We use $M/M/1$ queuing model to formulate this relation. Due to its simple form, $M/M/1$ queue is widely used in the operations management literature to model firm's capacity or pricing decisions when firms compete on time (e.g., Chen and Wan (2003, 2005); Hassin and Haviv (2003); Guo and Zhang (2010)). To simplify the analysis, we assume that the service rate of the public health system is fixed. This assumption is reasonable if the capacity or the productivity of the public health system is not easy

health system.

to adjust in the short run. In such a case, financial incentives could be an effective measure to influence the demand for public care. Nevertheless, we also discuss cases when the service rate is considered as a decision variable in Section 4.4 and Section 4.5. The assumption of exponential service time distribution of $M/M/1$ queue can be unrealistic in many cases. Therefore, we supplement the analytical results of $M/M/1$ queue with numerical analysis of $M/G/1$ queue, which assumes a general service time distribution.

The remainder of this chapter is organized as follows. Section 4.2 sets up the model. Section 4.3 to Section 4.5 are dedicated to presenting the analysis and results for the case when unequal welfare weights are assigned to patients. Section 4.3 discusses the optimal tax or subsidy decision to private care given a fixed public service rate. In Section 4.3, we first discuss the existence and uniqueness of first-order condition solutions, and then present the comparative statics of the optimal tax or subsidy with respect to welfare weight function and the public service rate. This is followed by the discussion of the improvement of income redistribution by implementing the optimal tax or subsidy. Section 4.4 discusses the decision of public service rate when the tax or subsidy on private care is given. Section 4.5 explores the optimization problem when the health planner makes tax/subsidy decision and public service rate decision jointly. Section 4.6 collects the results for the case when patients are treated equally. Section 4.7 provides concluding remarks.

4.2 Problem Formulation

We model the public health system as a single service station that has Poisson arrivals and exponentially distributed service time. Let λ be the effective arrival rate to the public health system and μ be the fixed service rate. The service rate here is defined as the rate of processing job requests. According to the results of $M/M/1$ queue, the expected waiting (queuing) time of patients in the public health system is $T = \left(\frac{1}{\mu-\lambda} - \frac{1}{\mu}\right)$. Similar to the assumption of health economics literature (e.g., Cullis and Jones (1985); Iversen (1993, 1997); Hoel and Sáther (2003); González (2005)), we assume zero waiting time in the private sector.

The provision of public care incurs two types of cost: production cost and capacity cost. The production cost has a constant marginal rate of q, so the total production cost is $q \cdot \lambda$; The capacity cost has a constant marginal constant rate c, so the total capacity cost is $c \cdot \mu$. For instance, the production cost can include the fee-for-service paid to physicians and the capacity cost can include the cost of maintaining hospital facilities or overhead cost. Let h be the unit price of private care and R be the health benefit of receiving one unit of treatment. We assume that the market for private care is competitive so that private health providers take price as given. We assume $h \geq q$.

The public health planner implements a monetary term t on each unit of private care. A patient needs to pay the consumer price of $h+t$ for one unit of private care. If $t > 0$, this monetary term serves as a tax levied on the consumption of private care. If $t < 0$, this monetary term is the public subsidy to private care. Let $p = h+t$, then p is the actual price that a patient needs to pay for private care.

4.2.1 Patient's Problem

Without loss of generality, we normalize the total patient arrival rate to be 1. In the rest of this chapter, we use **1** to represent the total patient arrival rate. Patients are heterogenous with respect to the cost of time in waiting, denoted by θ. We assume that θ follows a uniform distribution in $[0,1]$. A patient with time cost θ in the public health system has an expected net benefit of $(R - \theta T)$. The same patient has a net benefit of $(R-p)$ if he seeks private care. Each patient decides between public and private care by comparing the expected net benefits. To avoid triviality, we assume $R - h \geq 0$.

The patient type who is indifferent between public and private care, denoted by $\widetilde{\theta}$, must satisfy $R - \widetilde{\theta}T = R - p$. Therefore, we have $\widetilde{\theta} \cdot T = p$. Patients with time cost $\theta \leq \widetilde{\theta}$ will seek public care, while patients with time cost $\theta > \widetilde{\theta}$ will seek private care. The effective arrival rate to the public health system is:

$$\lambda = \int_0^{\widetilde{\theta}} 1 d\theta = \widetilde{\theta}$$

By $\widetilde{\theta}T = \widetilde{\theta}\left(\frac{1}{\mu-\lambda} - \frac{1}{\mu}\right) = p$, we must have

$$T = \frac{p}{\lambda} \quad \Rightarrow \quad \lambda = \alpha(p) \cdot \mu$$

where $\alpha(p) = \frac{1}{2}\left(\sqrt{p^2 + 4p} - p\right)$. We can see that $\alpha'_p > 0$ and $\alpha''_{pp} < 0$. For ease of exposition, we use λ to denote both the arrival rate to public health system and the patient type that is indifferent between public and private care.

For a patient with time cost θ, the expected cost that she needs to bear consists of two parts. The first part is the patient's share of the total cost of the public health system. We assume that the public health system is financed by a head tax (Besley and Coate, 1991), then the first part of cost is equal to $\frac{q\lambda+c\mu-t(1-\lambda)}{\mathbf{1}}$, where $q\lambda$ is the production cost, $c\mu$ is the cost of service rate, $t(1-\lambda)$ is the total amount of tax revenue collected from the private sector ($t>0$) or the total amount of subsidy provided to private care ($t<0$). The second part of cost is the expected cost of service, which is the cost of waiting $(=\theta T)$ if the patient chooses public care or the price p if he chooses private care. Let $B(\theta)$ be the expected cost of a patient with time cost θ, then we must have

$$B(\theta)=\frac{q\lambda+c\mu-t(1-\lambda)}{\mathbf{1}}+\min\{\theta T,p\}$$

Finally the welfare of a patient with time cost θ is defined as the health benefit net of the expected cost, i.e.,

$$R-B(\theta)$$

4.2.2 Health Planner's Problem

To consider the redistributive purpose, the health planner assigns unequal weights to the welfare of different patient types. We assume that patients with higher earning abilities have higher θ's. Since the health planner is more concerned about the welfare of persons with low earning abilities, she assigns proportionally higher weights to persons with lower time costs. We use a weight distribution function $w(\theta)$ to reflect this consideration. We assume that $w(\theta)$ decreases in θ. Without loss of generality, the sum of weights is normalized to 1, i.e., $\int_0^1 w(\theta)d\theta=1$. The weighted sum of patient welfare is:

$$\int_0^1 [R-B(\theta)]w(\theta)d\theta=R-\int_0^1 B(\theta)w(\theta)d\theta$$

Letting $V=\int_0^1 B(\theta)w(\theta)d\theta$, then maximizing the weighted sum of patient welfare is equivalent to minimizing the weighted sum of patient costs V:

$$V(t|\mu,w) = \underbrace{q\lambda + c\mu}_{\zeta_1(\mathrm{t}|\mu,\mathrm{w})} - \underbrace{t(1-\lambda)}_{\zeta_2(\mathrm{t}|\mu,\mathrm{w})} + \underbrace{T\int_0^\lambda \theta w(\theta)d\theta}_{\zeta_3(\mathrm{t}|\mu,\mathrm{w})} + \underbrace{p\int_\lambda^1 w(\theta)d\theta}_{\zeta_4(\mathrm{t}|\mu,\mathrm{w})} \tag{4.1}$$

$V(t|\mu,w)$ can be further partitioned into four parts of cost. Among them, $\zeta_1(t|\mu,w)$ is the cost of production and capacity of the public health system; $\zeta_2(t|\mu,w)$ is the amount of tax revenue received from the private sector or the amount of subsidy provided to private care; $\zeta_3(t|\mu,w)$ is the cost of waiting in the public system; $\zeta_4(t|\mu,w)$ is the amount of money that patients pay for private care. The sum of ζ_1 and $-\zeta_2$ is the amount of cost incurred in the public health system, so we call it "supply-side cost". The sum of ζ_3 and ζ_4 is the amount of cost that patients need to bear, so we call it "patient-side cost". The health planner's problem is:

$$\begin{aligned} &\min_t V(t|\mu,w) \\ s.t.\quad &\lambda = \alpha(p)\cdot\mu \\ &p = h+t \\ &0 \le \lambda \le 1 \end{aligned}$$

Since $p = h+t$, we can use p as the decision variable in the rest of this chapter. To make $0 \le \lambda \le 1$, we must have $p \in \left[0, \min\left\{R, \left[\frac{1}{\mu-1} - \frac{1}{\mu}\right]^+\right\}\right]$ (see Appendix C.1.1).

In the remainder of this chapter, the discussion of the health planner's optimization problem is based on the discussion of the first order condition (FOC) solution. Therefore, it behooves us to explicitly write out the derivative of V with respect to p. Substituting $\lambda = \alpha(p)\cdot\mu$ into (4.1) and taking the partial derivative of V with respect to p gives rise to:

$$\frac{\partial V}{\partial p} = \int_0^\lambda \left[\frac{\theta}{\lambda} - \theta\cdot\frac{p}{\lambda^2}\cdot\frac{\partial\lambda}{\partial p} - 1\right]\cdot w(\theta)d\theta + \lambda + (p-h+q)\cdot\frac{\partial\lambda}{\partial p} \tag{4.2}$$

where $\frac{\partial\lambda}{\partial p} = \alpha_p' \cdot \mu$ and $\alpha_p' = \left(\sqrt{1+\frac{4}{p^2+4p}} - 1\right)/2 > 0$. In the next two sections, we characterize the existence and uniqueness of FOC solutions. We then discuss how the optimal p determined by FOC solution responds to the weight function and the public service rate. It is followed by the discussion of the improvement of income redistribution by using the optimal tax or subsidy. Proofs of the propositions are in Appendix C.1.

4.3 Optimal Tax or Subsidy on Private Care

In this section, we study the optimal tax or subsidy decision. We first characterize the existence and uniqueness of the optimal tax or subsidy and then discuss how the optimal tax/subsidy respond to the changes of $w(\theta)$ and μ respectively. Next, we discuss how the improvement of income redistribution using the optimal tax/subsidy varies according to the public service rate. In order to obtain analytical results, we assume the following properties for the welfare weight function $w(\theta)$: $\lim\limits_{\theta\to 0_+} \theta \cdot w(\theta) = 0$. For example, both function $w(\theta) = (1-\sigma)\theta^{-\sigma}$, $\sigma \in (0,1)$ and piecewise weight function satisfy this assumption.

4.3.1 Existence and Uniqueness of FOC Solutions

To analyze the existence of FOC solutions, we study the functional properties of $\frac{\partial V}{\partial p}$ and we have the following results:

Proposition 4.1 *If patients are assigned unequal welfare weights, then we have*

(i) $\frac{\partial V}{\partial p}|_{p\to 0_+} \begin{cases} < 0, & h > q; \\ = 0, & h = q; \end{cases}$

(ii) $\frac{\partial V}{\partial p}|_{p\to p_{\max}} > 0;$

(iii) *The optimal tax/subsidy t^* is determined by FOC.*

The interpretation of Case 1 of Proposition 4.1 is as follows. If the price of private care is greater than the marginal production cost of public care, it not optimal for the health planner to fully subsidize private care. If the price of private care is equal to the marginal production cost of public care, it may not be optimal for the health planner to fully subsidizes private care either. This result is in contrast to the result of the equal welfare weight case that the health planner should always subsidize private care (to be shown in Section 4.6.1). This is because if the health planner fully subsidize private care, the total amount of subsidy would be borne equally by all patients, then it is equivalent to have each patient paying $h = q$ out of their pockets for private care. Instead, by not fully subsidizing private care, the health planner can induce only part of the patient population to the private sector, then the amount of supply-side cost would be smaller than $\mathbf{1} \cdot h$. Since the supply-side cost is shared equally among all patients, the amount of supply-side cost per capita is less than q for patients with low time costs. Since the health planner is more concerned about the welfare of patients with lower time costs, setting $p > 0$ (i.e., $t > -h$) results in better income redistribution. Case 2 of Proposition 4.1 states that the marginal total cost is always positive at the maximum net price. Case 3 of Proposition 4.1 states that the optimal p^* is determined by FOC solutions. This result agrees Proposition 7 of Hoel and Sáther (2003). The uniqueness of FOC solutions is hard to establish, but we can show that uniqueness is guaranteed for some meaningful weight functions. These weight functions would be used in the analysis in the later sections.

Proposition 4.2 *If the weight function $w(\theta)$ satisfies either of the following two conditions, then the FOC solution is unique.*

(i) *Both $w(\lambda)$ and $\frac{1}{\lambda^2}\int_0^\lambda \theta \cdot w(\theta) d\theta$ are convex in λ, and $\frac{\partial^2 V}{\partial p^2}|_{p \to 0_+} < 0$;*

(ii) $2 + \frac{2}{3}(h-q)\alpha^{-2} \geq \frac{2}{\lambda^2}\int_0^\lambda \theta \cdot w(\theta) d\theta$, *$\forall p \in R^+$.*

Case 1 of Proposition 4.2 guarantees that $\frac{\partial^2 V}{\partial p^2}$ is a quasi-convex function of p. Case 2 of Proposition 4.2 guarantees that $\frac{\partial^2 V}{\partial p^2}$ is a strictly increasing function of p. Either condition guarantees that $\frac{\partial V}{\partial p}$ is a unimodular function of p. Under either condition, $\frac{\partial V}{\partial p} = 0$ has at most one solution.

We show some examples of $w(\theta)$ that satisfy the conditions of Proposition 4.2. For Case 1 of Proposition 4.2, we have $h = q$ and $w(\theta) = a - (2a-2)\theta, a \in (1,2)$. This weight function has a constant rate $(2a-2)$ of decrease. For Case 2 of Proposition 4.2, any piecewise function $w(\theta)$ with $w(\theta) \leq 2$, $\forall \theta \in [0,1]$ would be qualified. For the specific weight function $w(\theta) = (1-\sigma)\theta^{-\sigma}, \sigma \in (0,1)$, the higher the σ, the proportionally higher the weights the health planner assigns to patients with lower time costs. To make $w(\theta) = (1-\sigma)\theta^{-\sigma}, \sigma \in (0,1)$ satisfy Case 2 of Proposition 4.2, we only need to have $h - q \geq 10^{-2}$.

In the next three subsections, we assume that FOC solution is unique so that p^* is a function of $w(\theta)$ and μ.

4.3.2 Change of the Optimal Tax/Subsidy in Response to Weight Function

In this subsection, we study how the optimal tax/subsidy responds to the change of weight function $w(\theta)$. A weight function $w_1(\theta)$ is said to stochastically dominate another weight function $w_2(\theta)$ if $\int_0^{\lambda}[w_1(\theta)-w_2(\theta)]d\theta > 0$, $\forall \lambda \in [0,1]$. This means that when the health planner makes tax or subsidy decision according to $w_1(\theta)$, she concerns more about the welfare of patients with low time costs than she does according to $w_2(\theta)$. For instance, when a new ruling party is elected, the party's decision making may switch from $w_1(\theta)$ to $w_2(\theta)$. Given a fixed μ, let p_1^* and p_2^* be the FOC solutions with respect to $w_1(\theta)$ and $w_2(\theta)$ respectively, then we have the following proposition:

Proposition 4.3 *Assume that patients are assigned unequal weights, the FOC solution is unique and weight function $w_1(\theta)$ stochastically dominates weight function $w_2(\theta)$, then we have $p_1^* > p_2^*$.*

The intuition behind Proposition 4.3 is straightforward. If the health planner is more concerned with the welfare of patients with low time costs, p^* should be higher so that either a higher tax is levied on the consumption of private care or a lower subsidy is provided to private care. For instance, de Vericourt and Lobo (2009) show that an Indian nonprofit organization uses revenue from for-profit activities to subsidize its for-free activities. In the former case, more tax revenue would be collected from the private sector to finance the public health system. In the later case, a lower subsidy is provided to private care so that some public money is saved.

4.3.3 Change of the Optimal Tax/Subsidy in Response to Public Service Rate

We define the utilization of the public health system to be the ratio of total patient arrival rate to the public service rate, i.e., $\mathbf{1}/\mu$. Given a fixed total patient arrival rate, the higher the service rate, the lower the utilization. We are interested to know how the optimal tax/subsidy changes when the utilization decreases. The following proposition states the main conclusion:

Proposition 4.4 *Given that patients are assigned unequal weights and the optimal tax/subsidy is determined by a unique FOC solution, we have* $\frac{\partial p^*(\mu)}{\partial \mu} < 0$*, i.e., when the utilization of the public health system decreases, the health planner should implement a lower tax or a higher subsidy.*

The intuition behind Proposition 4.4 is not immediately evident. To investigate why this result holds, we use a specific weight function $w(\theta) = (1-\sigma)\theta^{-\sigma}, \sigma \in (0,1)$ to illustrate the underlying forces that drive the result. According to (4.1), $\frac{\partial V}{\partial p}$ can be broken down into four parts of cost as follows:

$$\frac{\partial V}{\partial p} = \underbrace{\frac{\partial \zeta_1}{\partial p} - \frac{\partial \zeta_2}{\partial p}}_{\text{Marginal Supply-Side Cost}} + \underbrace{\frac{\partial \zeta_3}{\partial p} + \frac{\partial \zeta_4}{\partial p}}_{\text{Marginal Patient-Side Cost}} \tag{4.3}$$

where

$$\frac{\partial \zeta_1}{\partial p} = \alpha' q \mu > 0 \qquad \frac{\partial \zeta_2}{\partial p} = 1 - [\alpha + (p-h)\alpha']\,\mu$$

$$\frac{\partial \zeta_3}{\partial p} = \frac{1-\sigma}{2-\sigma}[\alpha + (1-\sigma)p\alpha']\,\alpha^{-\sigma}\mu^{1-\sigma} > 0 \qquad \frac{\partial \zeta_4}{\partial p} = 1 - [\alpha + (1-\sigma)p\alpha']\,\alpha^{-\sigma}\mu^{1-\sigma}$$

Some interpretations follow. When p increases, an increasing number of patients stay in the public health system, so both the production cost and the utilization of the public health system increase, i.e., both the marginal cost of production $\frac{\partial \zeta_1}{\partial p}$ and the marginal cost of waiting $\frac{\partial \zeta_3}{\partial p}$ are positive. Both the sign of the marginal cost of tax/subsidy $\frac{\partial \zeta_2}{\partial p}$ and the sign of the marginal cost of private care $\frac{\partial \zeta_4}{\partial p}$ depend on p and μ. We group these four marginal costs into marginal supply-side cost and marginal patient-side cost as in (4.3). The marginal patient-side cost is always positive and decreasing in p. The marginal supply-side cost is increasing in p. We also have $\left\|\frac{\partial^2 \zeta_1}{\partial p^2} - \frac{\partial^2 \zeta_2}{\partial p^2}\right\| > \left\|\frac{\partial^2 \zeta_3}{\partial p^2} + \frac{\partial^2 \zeta_4}{\partial p^2}\right\|$, i.e., the magnitude of the marginal supply-side cost is increasing in a faster rate than the magnitude of the marginal patient-side cost.

Now suppose that the public service rate is μ_1 and FOC yields a unique solution p_1^*, then we must have $\frac{\partial V}{\partial p}|_{\mu_1,p_1^*} = 0$, $\left(\frac{\partial \zeta_1}{\partial p} - \frac{\partial \zeta_2}{\partial p}\right)|_{\mu_1,p_1^*} < 0$, $\left(\frac{\partial \zeta_3}{\partial p} + \frac{\partial \zeta_4}{\partial p}\right)|_{\mu_1,p_1^*} > 0$ and $\left\|\frac{\partial \zeta_1}{\partial p} - \frac{\partial \zeta_2}{\partial p}\right\|_{\mu_1,p_1^*} = \left\|\frac{\partial \zeta_3}{\partial p} + \frac{\partial \zeta_4}{\partial p}\right\|_{\mu_1,p_1^*}$. For the second order cross partial derivatives with respect to p and μ, we have $\left(\frac{\partial^2 \zeta_1}{\partial p \partial \mu} - \frac{\partial^2 \zeta_2}{\partial p \partial \mu}\right)|_{\mu_1,p_1^*} > 0$, $\left(\frac{\partial^2 \zeta_3}{\partial p \partial \mu} + \frac{\partial^2 \zeta_4}{\partial p \partial \mu}\right)|_{\mu_1,p_1^*} <$

0 and $\left\|\frac{\partial^2\zeta_1}{\partial p\partial\mu} - \frac{\partial^2\zeta_2}{\partial p\partial\mu}\right\|_{\mu_1,p_1^*} > \left\|\frac{\partial^2\zeta_3}{\partial p\partial\mu} + \frac{\partial^2\zeta_4}{\partial p\partial\mu}\right\|_{\mu_1,p_1^*}$. Therefore, we have $\frac{\partial^2 V}{\partial p\partial\mu}|_{\mu_1,p_1^*} > 0$. For any μ_2 such that $\mu_2 = \mu_1 + \Delta\mu$ where $\Delta\mu > 0$, we must have $\frac{\partial V}{\partial p}|_{\mu_2,p_1^*} = \frac{\partial V}{\partial p}|_{\mu_1,p_1^*} + \int_{\mu_1}^{\mu_2}\frac{\partial^2 V}{\partial p\partial\mu}|_{\mu_1,p_1^*}d\mu > 0$. Since $\left\|\frac{\partial^2\zeta_1}{\partial p^2} - \frac{\partial^2\zeta_2}{\partial p^2}\right\| > \left\|\frac{\partial^2\zeta_3}{\partial p^2} + \frac{\partial^2\zeta_4}{\partial p^2}\right\|$, in order to obtain a FOC solution p_2^* for μ_2, we must have $p_2^* < p_1^*$. For more details about this analysis, please see Appendix C.2.

The intuition behind the above analysis is as follows. Keeping price p_1^* intact, if the public service rate μ increases, the public waiting time would decrease so that more patients stay in the public health system. In this case, the patient-side cost would decrease, but the supply-side cost would increase. The increase of the marginal supply-side cost is more than offsetting the decrease of the marginal patient-side cost, so the marginal total cost is increasing in μ. In order to obtain a new FOC solution, the health planner has to offset the increase of the marginal supply-side cost by decreasing p_1^*, i.e., lowering the tax or increasing the subsidy. Some patients with high time costs are therefore induced to the private sector, and the supply-side cost would increase at a lower rate. In a background paper to Boadway et al. (1998), Boadway et al. (1997) establish that the use of a subsidy to the consumption of private goods may be desirable if it is optimal to induce the high-ability persons to “opt-out” of using publicly provided goods. Nevertheless, the model of Boadway et al. (1997) is constructed in the standard optimal income taxation framework. Their results rely on the relaxation of the participation constraint that ensures the high-ability households are better off by not “opting in”. In contrast, our study does not impose such a constraint.

4.3.4 Threshold of Public Service Rate for Subsidy

By Proposition 4.4, we see that given a weight function $w(\theta)$, there must exist a public service rate μ_w such that for $\mu \leq \mu_w$, the health planner should implement tax (i.e., $p^* > h$); For $\mu > \mu_w$, the heath planner should implement subsidy (i.e., $p^* < h$). If there are two weight functions $w_1(\theta)$ and $w_2(\theta)$ such that $w_1(\theta)$ stochastically dominates $w_2(\theta)$, then we have the following proposition:

Proposition 4.5 *If weight function $w_1(\theta)$ stochastically dominates weight function $w_2(\theta)$, we must have $\mu_{w_1} \geq \mu_{w_2}$.*

Proposition 4.5 suggests that if the health planner is more concerned about the welfare of patients with lower time costs, the threshold for subsidy should increase.

4.3.5 Numerical Analysis

Proposition 4.3 and Proposition 4.4 are derived analytically using $M/M/1$ queue. However, the assumption of exponential service time distribution of $M/M/1$ queue may not be realistic in many cases. To test the robustness of Proposition 4.3 and Proposition 4.4 in the case of general service time distribution, we conduct extensive numerical analysis using $M/G/1$ queue. We use weight function $w(\theta) = (1 - \sigma)\theta^{-\sigma}$, where σ reflects the health planner's concern of income redistribution: the higher the σ, the proportionally higher the weights assigned to patients with lower time costs. The expected waiting time of patients in an $M/G/1$ queue is $\frac{\lambda(v^2+\mu^{-2})}{2(1-\rho)}$ where v is the standard deviation of the service time distribution and $\rho = \lambda/\mu$ is the traffic intensity. We allow v, q, c and h to vary in a wide range of values. The results of Proposition 4.3 and Proposition 4.4 hold in all of the numerical examples. One of these numerical example is Table C.1 of Appendix C.3. This example uses parameters $q = 0.2, c = 0.2, h = 0.5$ and $v = 0.5$. The service rate (by row) varies from $\mu = 1.10$ to $\mu = 3.00$ and the weight parameter σ (by column) varies from $\sigma = 0.1$ to $\sigma = 0.9$. Inside the table, positive numbers mean tax and negative numbers mean subsidy. For instance, when $\mu = 1.10$ and $\sigma = 0.1$, the table shows that the health planner should subsidize the consumption of each unit of private care by 0.317 monetary terms. The numerical results are consistent with the conclusions of Proposition 4.3, 4.4 and 4.5.

4.3.6 Benefits of the Optimal Tax/Subsidy on Income Redistribution

In this subsection, we discuss the types of public health system that benefit more on income redistribution from the optimal tax/subsidy. Using the optimal tax/subsidy, the health planner can always achieve a no worse income redistribution because $t = 0$ is a feasible solution. Let $V(0|\mu)$ denote the total cost when no tax or subsidy is used and $V(t^*|\mu)$ denote the total cost when the optimal tax/subsidy is used, then $J(\mu) = V(0|\mu) - V(t^*|\mu)$ is the amount of income redistribution improvement. We are interested to know whether or not this amount of improvement depends on the public service rate μ. By (4.1), we have

$$J(\mu) = V(0|\mu) - V(t^*|\mu) = \left\{ (h-q)\left[\alpha^* - \alpha_h\right] + \frac{h \cdot \alpha_h - p^* \cdot \alpha^*}{2} \right\} \mu$$

where $\alpha^* = \alpha(p^*), \alpha_h = \alpha(h)$. In order to yield analytical results, we limit our discussion to weight functions $w(\theta) = (1-\sigma)\theta^{-\sigma}$. We then have the following proposition:

Proposition 4.6 *Given weight functions $w(\theta) = (1-\sigma)\theta^{-\sigma}$, $\sigma \in (0,1)$, the optimal tax/subsidy would achieve larger income redistribution improvement when the public health service rate is either low or high.*

As high public service rate means low utilization, Proposition 4.6 suggests that the optimal tax/subsidy benefits more on income redistribution when the utilization of public health system is either high or low. This result is in contrast to that of the case of equal welfare weights (to be shown in Section 4.6.1) where the improvement of income redistribution is monotonously increasing at a constant rate in μ.

4.4 Optimal Public Service Rate Given A Fixed Tax/Subsidy

In this section, we study the service rate decision when the tax or subsidy on private care is given. This is an appropriate setting where the long term tax or subsidy rate is fixed, but the capacity of the public health system is allowed to adjust. We assume that the tax/subsidy t (thus $p = h + t$) is given, and the health planner determines μ to minimize the total weighted cost. Since $\mu = \frac{\lambda}{\alpha}$ and α is fixed by t, to minimize the cost function V over μ is equivalent to minimize V over λ . In the following analysis, we use λ as the decision variable. Taking partial derivative of V with respect to λ gives rise to:

$$\frac{\partial V}{\partial \lambda} = \underbrace{\frac{c}{\alpha} + q}_{\frac{\partial \zeta_1}{\partial \lambda}} + \underbrace{(p-h)}_{\frac{\partial \zeta_2}{\partial \lambda}} \underbrace{- \frac{p}{\lambda^2} \int_0^\lambda \theta \cdot w(\theta) d\theta}_{\frac{\partial \zeta_3}{\partial \lambda} + \frac{\partial \zeta_4}{\partial \lambda}} \qquad (4.4)$$

The marginal patient-side cost $\left(\frac{\partial \zeta_3}{\partial \lambda} + \frac{\partial \zeta_4}{\partial \lambda}\right)$ is always negative. The marginal cost of production $\frac{\partial \zeta_1}{\partial \lambda}$ is positive, because increasing the public service rate reduces the public waiting time so that more patients want to stay in the public health system. Therefore, there would be more public care to be produced. The marginal cost of tax/subsidy is positive if tax is used (i.e., $\frac{\partial \zeta_2}{\partial \lambda} > 0$ if $p > h$). This is because increasing public service rate attracts patients to the public health system, so less tax revenue is collected from the private sector. The marginal cost of tax/subsidy is negative if subsidy is used (i.e., $\frac{\partial \zeta_2}{\partial \lambda} < 0$ if $p < h$). This is because less patients would choose private care when the pubic service rate increases, so less subsidy is provided to private patients. $\frac{\partial V}{\partial \lambda}$ is an increasing function of λ. By studying the sign of $\frac{\partial V}{\partial \lambda}$ at the boundary points $\lambda_{\text{min}} = 0_+$ and $\lambda_{\text{max}} = \mathbf{1}_-$, we can summarize the health planner's optimal strategy as follows:

Proposition 4.7 *Given that patients are assigned unequal weights and the tax/subsidy is given,*

(i) *If* $\left\{\frac{c}{\alpha} + q + (p-h) - p\frac{w(0_+)}{2}\right\} \geq 0$, *then* $\frac{\partial V}{\partial \lambda} \geq 0$, $\forall \lambda \in [0,1]$, *i.e., V is strictly increasing in μ, so the health planner should set $\mu^* = 0$ such that all patients choose private care.*

(ii) *If* $\left\{\frac{c}{\alpha} + q + (p-h) - p\int_0^1 \theta \cdot w(\theta) d\theta\right\} < 0$, *then* $\frac{\partial V}{\partial \lambda} < 0$, $\forall \lambda \in [0,1]$, *i.e., V is strictly decreasing in μ, so the health planner should set $\mu^* = \frac{\mathbf{1}}{\alpha}$ such that all patients stay in the public health system.*

(iii) *Otherwise, μ^* is uniquely determined by FOC solution and μ^* is an interior point.*

Case 1 of Proposition 4.7 is likely to prevail under one or more of the following circumstances: (1) the health planner assigns insufficient weights to patients with the lowest time costs, i.e., when $w(0_+)$ is sufficiently small; (2) the cost of public service rate c is high; (3) the production cost of public care q is high. If Case 1 of Proposition 4.7 prevails, having patients to be served in the private sector is more efficient than doing so in the public health system. Therefore, the health planner's best strategy is to push all patients to the private sector.

Since $\frac{c}{\alpha} + q + (p-h) - p\int_0^1 \theta \cdot w(\theta) d\theta = \frac{c}{\alpha} + q + t\left(1 - \int_0^1 \theta \cdot w(\theta) d\theta\right) - h\int_0^1 \theta \cdot w(\theta) d\theta$, Case 2 of Proposition 4.7 is likely to prevail under one or more of the following circumstances: (1) either the cost of public service rate c or the production cost of public care q or both of them are low; (2) private patients are heavily subsidized, i.e., $t \ll 0$; (3) the price of private care h is sufficiently high. If Case 2 prevails, the health planner can save money from not subsidizing private patients and save patient's cost of private care if more patients stay in the public health system. Therefore, the health planner has the incentive to increase the public service rate to an extent that all patients choose to stay in the public health system.

In Case 3 of Proposition 4.7, it is neither optimal to set $\mu = 0$ to push all patients to the private sector, nor optimal to set μ high enough to accommodate all patients in the public health system. The interior optimal point λ^* solves the following equation:

$$\frac{1}{\lambda^2}\int_0^{\lambda} w(\theta)\theta d\theta = 1 - \frac{h-q}{p} + \frac{c}{p \cdot \alpha}$$

In this case, even if $h = q$, it is not optimal to set $\lambda^* = 0$ (i.e., $\mu^* = 0$), which is in contrast to the conclusion of Proposition 4.9.

4.5 Joint Decisions of Tax/Subsidy and Public Service Rate

This section discusses the health planner's optimization problem when she needs to decide the tax/subsidy and the public service rate jointly. Since the optimal tax/-subsidy is a function of μ, i.e., $p^*(\mu)$, we can substitute $p^*(\mu)$ for p in V and then optimize over μ. In this case, the partial derivative of V with respect to μ is:

$$\frac{\partial V(p^*(\mu)|\mu,w)}{\partial \mu} = \frac{\partial V}{\partial p}|_{p^*(\mu)} \cdot \frac{\partial p^*(\mu)}{\partial \mu} + \frac{\partial V}{\partial \mu}|_{p^*(\mu)}$$

Since $\frac{\partial V}{\partial p}|_{p^*(\mu)} = 0$, we have

$$\frac{\partial V(p^*(\mu)|\mu,w)}{\partial \mu} = \frac{\partial V}{\partial \mu}|_{p^*(\mu)}$$

If the health planner assigns unequal weights to patients, then we have

$$\frac{\partial V}{\partial \mu}|_{p^*(\mu)} = c + \left\{ \left[p\left(1 - \frac{\int_0^{\lambda} \theta \cdot w(\theta) d\theta}{\lambda^2}\right) - (h-q) \right] \alpha \right\}|_{p^*(\mu)} \tag{4.5}$$

The structural property of (4.5) is not easy to determine under a general weight function. By limiting our discussion to weight function $w(\theta) = (1-\sigma)\theta^{-\sigma}$, (4.5) can be simplified as:

$$\frac{\partial V}{\partial \mu}|_{p^*(\mu)} = c + \left\{ \left[\frac{p\sigma - (h-q)}{1+(1-\sigma)p\frac{\alpha'}{\alpha}} \right] \alpha \right\} |_{p^*(\mu)} \tag{4.6}$$

When $h = q$, (4.6) is positive for any $\mu \geq 0$, so the health planner should set $\mu^* = 0$ and push all patients to the private sector. When $h > q$, $\frac{\partial V}{\partial \mu}|_{p^*(\mu)}$ is a quasi-convex function of p^* and thus is a quasi-concave function of μ. The health planner's best strategy could be to serve only part of the patient population in the public health system while leaving the rest to the private sector. This result is in contrast to that of the case of equal welfare weights (to be shown in Section 4.6.3) where the health planner's best strategy is either to serve all patients in the public health system or to push all patients to the private sector.

4.6 Results for the Case of Equal Welfare Weight

This section collects results for the case when the health planner treats all patients equally, i.e., $w(\theta) = 1$. Subsection 4.6.1 characterizes the optimal tax/subsidy when the public service rate is given. Subsection 4.6.2 discusses the optimal public service rate when the tax/subsidy is given. Subsection 4.6.3 investigates the joint decisions of tax/subsidy and public service rate.

4.6.1 Optimal Tax/Subsidy on Private Care Given A Fixed Public Service Rate

By $w(\theta) = 1$, we have

$$\begin{aligned}
\frac{\partial V}{\partial p} &= \frac{1}{2}\left(p \cdot \frac{\partial \lambda}{\partial p} + \lambda\right) - (h-q) \cdot \frac{\partial \lambda}{\partial p} \\
\frac{\partial^2 V}{\partial p^2} &= \frac{\partial \lambda}{\partial p} + \left(\frac{p}{2} - h + q\right) \frac{\partial^2 \lambda}{\partial p^2} > 0
\end{aligned}$$

Therefore, the optimal net price p^* $(= h + t^*)$ is either at the boundary points, or determined by the unique FOC solution. We have the results for t^* and p^* as follows:

Proposition 4.8 *Given that the health planner treats all patients equally,*

(i) *If $h = q$, then $t^* = -q$ and $p^* = 0$.*

(ii) *If* $h > q$*, then* $-h < t^* < 0$ *and* $0 < p^* < h$.

(iii) $\frac{\partial t^*}{\partial h} < 0$ *and* $\frac{\partial p^*}{\partial h} > 0$.

(iv) $\frac{\partial t^*}{\partial \mu} = 0$.

Case 1 and 2 of Proposition 4.8 suggest that when patients are treated equally, the health planner should always provide subsidy to private care. Case 1 of Proposition 4.8 shows that if the price of private care is equal to the marginal production cost of public care, the health planner should provide full subsidy to patients so that all patients go to the private sector. This is exactly the same conclusion of Proposition 8 of Hoel and Sáther (2003). Case 2 of Proposition 4.8 shows that if the price of private care is greater than the marginal production cost, providing full subsidy is no longer optimal. Case 3 of Proposition 4.8 states that when the price of private care increases, the health planner should provide higher subsidy. However, the rate of subsidy increase should be lower than the rate of price increase. Case 4 of Proposition 4.8 states that the optimal subsidy is independent of the public service rate, which is in contrast to the results of the case of unequal welfare weights.

Similar to the discussion of unequal welfare weights, we use $V(0|\mu)$ to denote the total cost when no subsidy is used and use $V(t^*|\mu)$ to denote the total cost when the optimal subsidy is used. Therefore, $J(\mu) = V(0|\mu) - V(t^*|\mu)$ is the improvement of income redistribution by using the optimal subsidy. By (4.1), we have

$$J(\mu) = V(0|\mu) - V(t^*|\mu) = \left\{ (h-q)\left[\alpha^* - \alpha_h\right] + \frac{h \cdot \alpha_h - p^* \cdot \alpha^*}{2} \right\} \mu$$

where $\alpha^* = \alpha(p^*), \alpha_h = \alpha(h)$. The marginal income redistribution improvement with respect to μ is:

$$\frac{\partial J}{\partial \mu} = (h-q)\left[\alpha^* - \alpha_h\right] + \frac{h \cdot \alpha_h - p^* \cdot \alpha^*}{2} > 0$$

The improvement of income redistribution is monotonously increasing in μ at a constant rate. Therefore, the higher the public service rate, the more the optimal subsidy improves income redistribution. Again, this conclusion is in contrast to the results of the case of unequal welfare weights.

4.6.2 Optimal Public Service Rate Given A Fixed Tax/Subsidy

If $w(\theta) = 1$ and t is given, according to (4.1), the total cost V becomes

$$V(\mu|t) = q + c\mu + (h-q)(1-\lambda) + \frac{p\lambda}{2}$$

The first-order derivative with respect to μ is

$$\frac{\partial V}{\partial \mu} = c + \left[\frac{p}{2} - (h-q)\right]\alpha \tag{4.7}$$

Note that the right hand side of (4.7) is independent of μ. Depending on the sign of the right hand side, the cost function V is either strictly decreasing or strictly increasing in μ. Therefore, we have the following proposition:

Proposition 4.9 *Given that the tax/subsidy is fixed and all patients are treated equally,*

(i) *If* $c + \left[\frac{p}{2} - (h-q)\right]\alpha > 0$*, then the health planner should set* $\mu^* = 0$ *so that* $\lambda^* = 0$.

(ii) *Otherwise, the health planner should set* $\mu^* = \frac{\boldsymbol{1}}{\alpha}$ *such that* $\lambda^* = \boldsymbol{1}$.

If Case 1 of Proposition 4.9 prevails, the health planner should push all patients to the private sector; Otherwise, the health planner should set the public service rate high enough to accommodate all patients in the public health system. We can see that $c + \left[\frac{p}{2} - (h-q)\right]\alpha > 0$ is likely to hold when c is high or h is low. In this case, providing public care to patients is less efficient than having them served in the private sector, so the health planner should push all patients to the private sector. By setting $\mu = 0$, patients have no choice but to accept private care. The total amount of tax revenue collected from the private sector or the total amount of subsidy provided to private care is equal to $(t \cdot \mathbf{1})$. This amount of tax revenue or subsidy would be paid back to patients, so at the end each patient only bears a cost of h.

4.6.3 Joint Decisions of Tax/Subsidy and Public Service Rate

Section 4.6.1 shows that if p^* is determined by FOC solution, then p^* is independent of μ but strictly increasing in h. Therefore, we can write p^* as a function of h, i.e., $p^*(h)$. Substituting $p^*(h)$ for p in V and taking the first order derivative of μ yields:

$$\frac{\partial V}{\partial \mu}|_{p^*(\mu)} = c - \frac{\alpha^2}{2\alpha'}|_{p^*(h)}$$

Depending on the sign of $c - \frac{\alpha^2}{2\alpha'}|_{p^*(h)}$, the health planner either sets $\mu^* = 0$ or sets μ^* high enough to accommodate all patients in the public health system. Because $c - \frac{\alpha^2}{2\alpha'}|_{p^*(h)}$ is a strictly decreasing function of h, the health planner's optimal strategies can be summarized as the following proposition:

Proposition 4.10 *Let $\widetilde{h}$ be the solution to equation $c - \frac{\alpha^2}{2\alpha'}|_{p^*(h)} = 0$, then*

(i) *If $h \leq \widetilde{h}$, then $\mu^* = 0$, so $\lambda^* = 0$ and $V = h$.*

(ii) *If $h > \widetilde{h}$, μ^* must solve $\frac{1}{\mu-1} - \frac{1}{\mu} = p^*(h)$, so $\lambda^* =$ **1** and $V = c\mu^* + q + \frac{p^*(h)}{2}$.*

Proposition 4.10 suggests that if the health planner treats all patients equally and the price of private care is lower than the threshold $\widetilde{h}$, the health planner should push all patients to the private sector; Otherwise, if the price of private care is high for patients, the health planner should accommodate all patients in the public health system.

4.7 Concluding Remarks

Public provision of private goods, such as public health care, may serve redistributive purpose even if the public provision is financed by a head tax. Many OECD countries subsidize private care to reduce the demand for public care and thus reduce the long public waiting time. In this chapter, we showed that subsidy to private care can also be used to improve income redistribution under certain circumstances. In the case of unequal welfare weights, the health planner should subsidize private care when the utilization of the public health system is low. This is because the subsidy would induce patients with higher time costs to the private sector, so the demand for public care would fall and the head tax that finances the public provision is reduced. As shown in Section 4.3.3, this result arises as a consequence of the health planner balancing the competing supply-side cost and patient-side cost. In the case of equal welfare weights, the health planner should always subsidize private care. We also discussed the choice of public service rate when the tax or subsidy on private care is given. We found that in most cases, the health planner should either set the public service rate to zero to push all patients to the private sector, or set the public service rate high enough to serve all patients. The decision depends on the efficiency of health care production in each sector.

Chapter 5

Conclusion

5.1 Summary

Whether or not the existence of private care would reduce public health waiting times and improve social welfare has long been a central topic for research in health economics (Cullis and Jones, 1985). Due to the complexity of health care system, the answer to this question depends on the underlying institutional setting that a study is based on. Motivated by the current public versus private debates in Canada's health reform debates, this thesis uses both theoretical and empirical methodologies to investigate some of the core issues raised in the debates: Chapter 2 and Chapter 4 investigate the issue of private care financing; Chapter 3 investigates the issue of physician dual practice. This thesis covers both dimensions of research on private care in health economics. In particular, the three essays included in this thesis are either of empirical nature or motivated by empirical evidence: Chapter 2 is an empirical analysis, Chapter 3 is a theoretical research motivated by the empirical evidence of Manitoba's cataract surgeries, and Chapter 4 is a theoretical research motivated by the empirical findings of Chapter 2. This thesis contributes to the existing literature of health economics on private care by providing managerial insights to some controversial issues raised in the public versus private debates based on Canada's institutional setting.

Chapter 2 empirically investigates the impact of allowing private care financing on public waiting times. Using joint replacement data of nine Canadian provinces, we test two policies that induce private care financing. Due to the limitations of the data set, we rely on cross sectional analysis and Random-effects model to derive the results and test the robustness of these results. Given the available data to our study and the methodological approaches we employ, the results show that the two policies are associated with shorter waiting times. In particular, the policy of providing subsidy to patients seeking private care is consistently significant in all regressions. The contribution of this study to the existing literature is to provide an empirical analysis of private care financing and public waiting time under Canada's institutional setting.

Chapter 3 investigates the impact of allowing physician dual practice on public waiting time and patient welfare. This topic is relevant as physician dual practice is currently prohibited in all Canadian provinces, but proposals to approve it had been made in the health reform debates. By examining the empirical evidence of Manitoba's cataract surgeries in the late 1990's when physician dual practice was allowed, we formulate a queuing model to study the system performance and patient welfare with and without physician dual practice. Two effects induced by allowing physician dual practice are considered in the model: one is the quality differentiation between dual-practice physicians and public-only physicians, and the other is the prioritization of patients in the queue of dual-practice physicians. We find that allowing physician dual practice results in a longer waiting time for patients with lower time costs due to the prioritization effect. However, these patients might benefit from allowing physician dual practice because the quality differentiation effect allows them to self-select the service of a higher quality in the public health system. The existing literature on physician dual practice is recent and very limited, so our study has provided new insights to this topic.

Chapter 4 is related to the findings of Chapter 2. Chapter 2 shows that providing subsidy to private care may contribute to waiting time reduction. This study addresses a related and relevant question: in addition to waiting time reduction, whether or not providing subsidy to private care would also improve income redistribution. This is a research question that has never been investigated by the existing literature of public provision of private goods. The stylized model used in our study assumes that the operation of the public health system is financed by a head tax. We find that providing subsidy to private care can improve income redistribution by inducing patients with higher earning abilities to the private sector, so the production cost of public care would be reduced and so would be the head tax that everyone, including patients with lower earning abilities, pays for. The conditions under which providing subsidy to private care can improve income redistribution are derived.

5.2 Related Ongoing Work

5.2.1 Efficiency of Different Subsidy Schemes on Reducing Public Waiting Time and Improving Patient Welfare

This work is related to both Chapter 2 and Chapter 4. Under the pressure of reducing long public waiting times, governments are using different schemes to subsidize patients who seek private care in order to take the demand burden off the public health system. For instance, Quebec government introduces a waiting time guarantee policy for cataract and joint replacement surgeries: If the treatment cannot be provided to a patient within the guaranteed period of nine months, the health ministry of Quebec must purchase the treatment from a heath provider who operates outside Quebec's public health system.[1] In addition to Quebec, United Kingdom and Sweden also implement similar waiting time guarantee policies. Under these policies, patients are fully subsidized to receive private treatments should their waiting times exceed the guarantee. For brevity, we call it a "full subsidy" scheme. A full subsidy scheme works in a public-private mixed health system.

In contrast to the full subsidy scheme, another mechanism used to reduce public waiting times is to provide partial subsidies to patients seeking private care. Such a device can alleviate the congestion of the public health system by rationing patients into the private sector. Examples of governments implementing this scheme have been introduced in Chapter 2. Empirical studies of Siciliani and Hurst (2005) and Chapter 2 show that this partial subsidy policy contributes to the waiting time reduction. Additionally, Chapter 4 shows that, under certain conditions, providing partial subsidies to patients seeking private care improves social welfare.

Which subsidy mechanism is more efficient in achieving a larger amount of patient welfare given the same amount of public health budget? This is an important question to the policy makers. At first, the answer seems to be blatantly obvious. To be cost-efficient, subsidies shall only be applied to those patients who join the private sector but would join the public system without subsidy. The full subsidy scheme

[1] See "Guaranteeing Access: Meeting the challenges of equity, efficiency and quality". Retrieved from Gouvernement du Québec website: http://publications.msss.gouv.qc.ca/acrobat/f/documentation/2005/05-721-01A.pdf

indeed can achieve that since only those patients who are routed from the public to the private system receive the subsidy. On the contrary, patients who would not join the public system even without subsidy are subsidized under the partial subsidy scheme. On a second thought, a “how” question looms: How could a mechanism ensure the allocation of the subsidy to those in need? The brutal fact confronted by those patients under the full subsidy scheme is that: to be subsidized, they must wait. Clearly, when waiting plays a rationing role, a big waste of social welfare occurs. In contrast, partial subsidy scheme avoids such a rationing cost by providing subsidy unconditionally to those seeking private care. Therefore, although the the full subsidy scheme has the advantage of “allocation efficiency”, it also incurs a “rationing cost”. Consequently, we shall invest more scrutiny to examine the pros and cons of the two subsidy schemes.

In Guo and Qian (2011b), we set up mathematical models to evaluate the efficiency of the two subsidy schemes in reducing public waiting times and improving patient welfare. The congestion in the public system is modeled through a queueing model. The central point is to model patient’s strategic choice behavior, given certain levels of delay information and the conditions of obtaining the subsidy. We consider the private care as an outside option for patients. Patients who seek private care are “balking” patients, so a patient arriving to the public system needs to decide whether “to join” the public system or “to balk”. When making such decisions, a patient must take others’ decisions into consideration because delay is endogenous. We model the equilibrium behavior of patients. We show that equilibria could be different should different levels of delay information be provided to patients. We model different levels of delay information by considering the queue to be either unobservable or observable. In the case of unobservable queue, patients can only form an expectation of the congestion given others’ decisions. In the cases of observable queue, patients can be provided information of actual queue length or actual workload upon arrival. Armored with the equilibrium analysis on patient behavior, we then consider the optimal design of the two mechanisms. In the partial subsidy scheme, the decision variable is the amount of subsidy per patient whereas, in the full subsidy scheme, the decision variable is the threshold level of waiting time for a patient to be qualified to switch to the private sector.

The existing literature on strategic customer behavior and queuing is abundant. Noar (1969) pioneers the study on the strategic customer behavior and socially optimal control with observable queue. Rich work is carried over ever since then and the literature dated before year 2003 is surveyed by Hassin and Haviv (2003). Two important streams of this field emerge in recent years. The first stream is the study of the impact of different levels of delay information on the system performance, e.g., Hassin (1986); Guo and Zipkin (2007, 2008, 2009). One major finding of this literature is that providing more accurate delay information needs not necessarily to improve customer's utility and service provider's profit. Other literature studying delay information and customer behavior includes Armony and Maglaras (2004); Mandelbaum and Shimkin (2000); Shimkin and Mandelbaum (2004); Allon et al. (2008); Allon and Bassamboo (2008); Ibrahim and Whitt (2009a,b); Armony et al. (2009); Jouini et al. (2009). Besides this stream, there also exists an increasing trend on studying customer strategic behavior in queues where positive externalities and follow-the-crowd (FTC) behavior are observed, e.g., Burnetas and Economou (2007); Economou and Kanta (2008); Economou et al. (2011); Guo and Hassin (2011); Johari and Kumar (2008); Veeraraghavan and Debo (2009). In the following paragraphs, the contributions of our work to the existing literature and to the managerial insights are discussed.

The equilibrium analysis parts under the partial subsidy scheme with unobservable and observable queues are, de facto, the same as Noar (1969) and Edelson and Hildebrand (1975), respectively. However, the optimal control problem is different: in those papers, the pricing decision is unconstrained; whereas it needs to satisfy the budget constraint in our case. This difference breeds an interesting finding: under the unobservable queue, the socially optimal solution must be a boundary solution with the budget constraint binding; whereas under the observable queue, the socially optimal solution can be an interior solution with the budget constraint unbinding. This is counter-intuitive: one may believe that the more the public fund is spent, the larger the social welfare. Our findings show that, when the delay information is revealed to customers, a naive implementation of the spending-all-budget policy could actually reduce social welfare.

The equilibrium analysis under the full subsidy scheme is analogous to the one under the partial subsidy scheme except that there exist both balking and "reneging" (being switched to the private sector with full subsidy) behaviors: when patients decide to join or balk upon arrival, they need to consider the later chance of reneging with full subsidy. Such a system has a more complex expression of system performance measures and in general needs numerical calculations. Despite these complexities, we are able to carry out the optimization over the optimal threshold level and show that it still holds that the socially optimal solution might not satisfy a binding budget constraint under observable queue.

A comparison over patient welfare under the two subsidy schemes yields the main conclusion. Analytically, we show that the partial subsidy scheme is better than the full subsidy scheme with a scarce fund, while the full subsidy scheme prevails with an ample fund. In the case of moderate fund, numerical studies reveal that there exists a critical level, below which the partial subsidy scheme is better, while above which the full subsidy scheme prevails. We also find a sufficient condition for this conclusion to hold. This conclusion is reasonable and reflects the relative roles of "rationing cost" versus "allocation efficiency" along with the total fund: when the total subsidy fund is large, the threshold for rationing can be set to be relatively low, which diminishes the marginal rationing cost of the full subsidy scheme. Therefore, the advantage of the full subsidy scheme looms with a large fund.

5.2.2 Subsidy on Private Care versus Public Capacity Expansion

Section 5.2.1 considers the scenario when the public health planner can only choose among different subsidy schemes to reduce public waiting times. This scenario is relevant when the public health system cannot be easily expanded in the short run, so providing financial incentives becomes an effective measure to influence patient's behavior. From a long term perspective, the more important question is whether the limited health budget should be used to expand the capacity of the public system or to subsidize private care. In (Guo and Qian, 2011a), we address this related and important question. In this work, we assume that the health planner can split the health budget to serve both purposes. We aim to answer the question as under what circumstances, providing subsidy to private care is more efficient to improve patient's welfare than expanding the capacity of public system.

Due to analytical difficulties, Guo and Qian (2011a) only consider homogeneous patients. The problem formulation is similar to that of Section 4.6.3, i.e., the case of equal welfare weights. However, the optimization problem in Section 4.6.3 is unconstrained; while the joint decisions of capacity and subsidy need to satisfy the budget constraint in Guo and Qian (2011a). The introduction of a budget constraint into the problem makes Part 2 of Proposition 4.10 no longer hold. Our results show that the health planner should always use some part of the budget, if not all of it, to subsidize patients seeking private care. However, there exist no conditions under which the budget should be solely used to expand the capacity of the public system.

For the sake of simplification, a competitive private care market is assumed in Guo and Qian (2011a). However, our model can be easily extended to other market structures, under which the behavior of the private care providers needs to be considered. On the one hand, providing subsidy to private care may encourage the private health providers to raise the price. On the other hand, providing subsidy to private care may also attract more health providers to the private care market, which would intensify the price competition. Therefore, it would be of our interests to see how these market structures would affect the public health planner's decision making.

5.3 Future Research Plan

5.3.1 Empirical Study of CJRR Data in A Longer Time Horizon

Although the empirical study of Chapter 2 provides supportive evidence to one side of the "public versus private" debates, we should not take the result as conclusive due to data limitation. In particular, the data truncation problem only allows us to generate a small sample for test. This data limitation suggests the need for more data collection and empirical studies of this subject. It would be useful if the same analysis can be applied to more waiting time data in a longer time horizon to test the robustness of our findings. In order to do this, we are in the process of applying for CJRR data in the most recent years (2007-2010). With the addition of new data, we would be able to extend the current study to examine the effects of policy changes in recent years. For instance, Quebec government changed its policies towards public waiting times and these policies took effect in January 2008. It will be interesting to see how these policy changes affect the public health waiting times. Additionally, more data may also allow us to empirically isolate the demand-side effect from the supply-side effect. We see these exercises as a natural extension of Chapter 2.

5.3.2 Empirical Study of Physician Dual Practice

The phenomenon of physician dual practice has triggered substantial interests in health economists and thus the theoretical studies of physician dual practice are growing rapidly. Nevertheless, the empirical side of the field has not yet been developed. We see it as a natural direction for future research. Due to the data accessibility policy of Manitoba government, we are still unable to conduct a related empirical study on Manitoba's cataract surgery data. The findings of empirical study, if successfully obtained, would provide us additional perspectives to look into the problem and generate more interesting questions for theoretical research.

Bibliography

Allon, G., A. Bassamboo. 2008. The impact of delaying the delay announcements. *Operations Research* forthcoming. → pages 85

Allon, G., A. Bassamboo, I. Gurvich. 2008. "We will be right with you": Managing customer expectations with vague promises and cheap talk. *Operations Research* forthcoming. → pages 85

Armony, M., C. Maglaras. 2004. Contact centers with a call-back option and real-time delay information. *Operations Research* **52**(4) 527–545. → pages 85

Armony, M., N. Shimkin, W. Whitt. 2009. The impact of delay announcements in many-server queues with abandonment. *Operations Research* **57**(1) 66–81. → pages 85

Besley, T., S. Coate. 1991. Public provision of private goods and the redistribution of income. *American Economic Review* **81**(4) 979–984. → pages 61, 65

Besley, T., J. Hall, I. Preston. 1999. The demand for private health insurance: Do waiting lists matter? *Journal of Public Economics* **72**(2) 155–181. → pages 2, 9

Biglaiser, G., C.A. Ma. 2007. Moonlighting: public service and private practice. *RAND Journal of Economics* **38**(4) 1113–1133. → pages 32

Bir, A., K. Eggleston. 2003. Physician dual practice: Access enhancement or demand inducement? Working paper, Department of Economics, Tufts University, Working Paper No.2033-11. → pages 32, 46, 57

Blendon, R.J., J.M. Benson. 2009. Understanding how Americans view health care reform. *The New England Journal of Medicine* **361**(9) e13. → pages 8

Blomquist, S., V. Christiansen. 1995. Public provision of private goods as a redistributive device in an optimum income tax model. *Scandinavian Journal of Economics* **97**(4) 547–567. → pages 61

Boadway, R., M. Marchand. 1995. The use of public expenditure for redistributive purposes. *Oxford Economic Papers New Series* **47**(1) 45–59. → pages 61

Boadway, R., M. Marchand, M. Sato. 1997. Subsidies versus public provision of private goods as instruments for redistribution. Working paper 942, Queen's University, Kingston, Canada. → pages 71

Boadway, R., M. Marchand, M. Sato. 1998. Subsidies versus public provision of private goods as instruments for redistribution. *Scandinavian Journal of Economics* **100**(3) 545–564. → pages 61, 62, 71

Bucovetsky, S. 1984. On the use of distributional waits. *Canadian Journal of Economics* **17**(4) 699–717. → pages 61

Burnetas, A., A. Economou. 2007. Equilibrium customer strategies in a single server markovian queue with setup times. *Queueing Systems* **56** 213–228. → pages 85

Canadian Institute for Health Information. 2008a. Canadian joint replacement registry (CJRR) 2007 annual report. Research report, Ottawa, Ontario, Canada. → pages 8, 18, 20, 27

Canadian Institute for Health Information. 2008b. National health expenditure trends, 1975-2008. Research report, Ottawa, Ontario, Canada. → pages 2, 25

Chen, H., Y.W. Wan. 2003. Price competition of make-to-order firms. *IIE Transactions* **35** 817–832. → pages 33, 62

Chen, H., Y.W. Wan. 2005. Capacity competition of make-to-order firms. *Operations Research Letters* **33** 187–194. → pages 33, 62

Cipriano, L.E., B.M. Chesworth, C.K. Anderson, G.S. Zaric. 2007. Predicting joint replacement waiting times. *Health Care Management Science* **10**(2) 195–215. → pages 9, 13

Conner-Spady, B., G. Johnston, J. McGurran, M. Kehler, T. Noseworthy. 2008. Willingness of patients to change surgeons for a shorter waiting time for joint arthroplasty. *Canadian Medical Association Journal* **179**(4) 327–332. → pages 34

Cremer, H., F. Gahvari. 1997. In-kind transfers, self-selection and optimal tax policy. *European Economic Review* **41** 97–114. → pages 61

Cullis, J.G., P.R. Jones. 1985. National Health Service waiting lists: A discussion of competing explanations. *Journal of Health Economics* **4** 119–135. → pages 16, 63, 81

de Vericourt, F., M. Lobo. 2009. Resource and revenue management in nonprofit operations. *Operations Research* **57**(5) 1114–1128. → pages 69

DeCoster, C., L. MacWilliam, R. Walld. 2000. Waiting times for surgery: 1997/8 and 1998/9 update. Research report, Manitoba Centre for Health Policy and Evaluation, Winnipeg, Manitoba, Canada. → pages 30, 31, 38, 48, 52

Derrett, S., T.H. Bevin, P. Herbison, C. Paul. 2009. Access to elective surgery in New Zealand: Considering equity and the private and public mix. *The International Journal of Health Planning and Management* **24**(2) 147–160. → pages 3

Duckett, S.J. 2005. Private care and pubic waiting. *Australian Health Review* **29**(1) 87–93. → pages 3

Economou, A., A. Gómez-Corral, S. Kanta. 2011. Optimal balking strategies in single-server queues with general service and vacation time. Working paper, University of Athens, Greece. → pages 85

Economou, A., S. Kanta. 2008. Equilibrium balking strategies in the observable single-server queue with breakdowns and repairs. *Operations Research Letters* **36** 696–699. → pages 85

Edelson, N., K. Hildebrand. 1975. Congestion tolls for poisson queuing processes. *Econometrica* **43** 81–92. → pages 85

Eggleston, K., A. Bir. 2006. Physician dual practice. *Health Policy* **78** 157–166. → pages 3, 5, 45, 59, 62

Esmail, N., M. Walker. 2005. Waiting your turn: Hospital waiting lists in Canada (15th ed). Research report, Fraser Institute, Vancouver, British Columbia, Canada. → pages 18

Esmail, N., M. Walker. 2006. Waiting your turn: Hospital waiting lists in Canada (16th ed). Research report, Fraser Institute, Vancouver, British Columbia, Canada. → pages 18

Esmail, N., M. Walker. 2008. How good is Canadian health care? 2008 report. Research report, Fraser Institute, Vancouver, British Columbia, Canada. → pages 2

Evans, R.G. 2000. Canada: How the system works. A summary. *Journal of Health Politics, Policy and Law* **25**(5) 889–897. → pages 10, 11

Flood, C., T. Archibald. 2001. The illegality of private health care in Canada. *Canadian Medical Association Journal* **164**(6) 852–830. → pages 2, 3, 10, 20, 26

García-Pardo, A., P. González. 2007. Policy and regulatory responses to dual practice in the health sector. *Health Policy* **84** 142–152. → pages 59

González, P. 2004. Should physicians' dual practice be limited? An incentive approach. *Health Economics* **13** 505–524. → pages 32, 45

González, P. 2005. On a policy of transferring public patients to private practice. *Health Economics* **14** 513–527. → pages 59, 63

Guo, P., R. Hassin. 2011. Strategic behavior and social optimization in markovian vacation queues. *Operations Research* **59** 986–997. → pages 85

Guo, P., Q. Qian. 2011a. Public capacity expansion versus private service subsidization in a two-tier service system. Working paper, Department of Logistics and Maritime Studies, Hong Kong Polytechnic University. → pages 86, 87

Guo, P., Q. Qian. 2011b. Utilizing subsidy schemes to reduce waiting times for public health care. Working paper, Department of Logistics and Maritime Studies, Hong Kong Polytechnic University. → pages 84

Guo, P., G. Zhang. 2010. Pricing, capacity and financing decisions in a two-tier service system. Working paper, Department of Logistics and Maritime Studies, Hong Kong Polytechnic University, Hong Kong, China. → pages 62

Guo, P., P. Zipkin. 2007. Analysis and comparison of queues with different levels of delay information. *Management Science* **53**(6) 962–970. → pages 85

Guo, P., P. Zipkin. 2008. The effects of information on a queue with balking and phase-type service times. *Naval Research of Logistics* **55** 406–411. → pages 85

Guo, P., P. Zipkin. 2009. The effect of the availability of waiting-time information on a balking queue. *European Journal of Operational Research* **198** 199–209. → pages 85

Hassin, R. 1986. Consumer information in markets with random products quality: The case of queues and balking. *Econometrica* **54** 1185–1195. → pages 85

Hassin, R., M. Haviv. 2003. *To Queue or Not to Queue: Equilibrium Behavior in Queuing Systems*. Kluwer, Boston, MA. → pages 33, 62, 85

Hoel, M., E. Sáther. 2003. Public health care with waiting time: The role of supplementary private health care. *Journal of Health Economics* **22** 599–616. → pages 6, 61, 62, 63, 68, 77

Hurley, J., R. Vaithianathana, T.F. Crossley, D. Cobb-Clark. 2001. Parallel private health insurance in Australia: A cautionary tale and lessons for Canada. Working paper, McMaster University Centre for Health Economics and Policy Analysis, Hamilton, Ontario, Canada. Working paper no.01-12. → pages 2

Ibrahim, R., W. Whitt. 2009a. Real-time delay estimation based on delay history. *Manufacturing & Service Operations Management* **11**(3) 397–415. → pages 85

Ibrahim, R., W. Whitt. 2009b. Real-time delay estimation in overloaded multiserver queues with abandonments. *Management Science* **55**(10) 1729–1742. → pages 85

Iversen, T. 1993. A theory of hospital waiting lists. *Journal of Health Economics* **12** 55–71. → pages 3, 63

Iversen, T. 1997. The effect of a private sector on the waiting time in National Health Service. *Journal of Health Economics* **16**(4) 381–396. → pages 3, 9, 16, 63

Jan, S., Y. Bian, M. Jumpa, Q. Meng, N. Nyazema, P. Prakongsai, A. Mills. 2005. Dual job holding by public sector health professionals in highly resource-constrained settings: Problem or solution? *Bulletin of the World Health Organization* **83** 771–776. → pages 59

Jofre-Bonet, M. 2000. Public health care and private insurance demand: The waiting time as a link. *Health Care Management Science* **3**(1) 51–71. → pages 9

Johari, R., S. Kumar. 2008. Externalities in services. Working paper, Graduate School of Business, Stanford University, Stanford, CA. → pages 85

Johnson, T., K. Morris, J. Zelmer. 2007. Priority areas for wait time reduction: What we know and what we don't know. *Healthcare Quarterly* **10**(3) 26–28. → pages 26

Jouini, O., Y. Dallery, O.Z. Aksin. 2009. Queuing models for multiclass call centers with real-time anticipated delays. *International Journal of Production Economics* **120** 389–399. → pages 85

Karlson, E.W., L.A. Mandl, G.N. Aweh, O. Sangha, M.H. Liang, F. Grodstein. 2003. Total hip replacement due to osteoarthritis: The importance of age, obesity, and other modifiable risk factors. *The American Journal of Medicine* **114**(2) 93–98. → pages 20

Kondro, W. 2007. Federal budget delivers on health care but still disappoints. *Canadian Medical Association Journal* **176**(8) 1071. → pages 2

Linsay, C.M., B. Feigenbaum. 1984. Rationing by waiting lists. *American Economic Review* **74**(3) 404–417. → pages 14

Mandelbaum, A., N. Shimkin. 2000. A model for rational abandonments from invisible queues. *Queueing Systems* **36** 141–173. → pages 85

Martin, S., P.C. Smith. 1999. Rationing by waiting lists: An empirical investigation. *Journal of Public Economics* **71**(1) 141–164. → pages 12

Martin, S., P.C. Smith. 2003. Using panel methods to model waiting times for National Health Service surgery. *Journal of the Royal Statistical Society* . → pages 24

McPake, B., L. Kumaranayake, C. Normand. 2002. *Health Economics: An International Perspective*. 1st ed. Routledge, London, England. → pages 60

Mirrlees, J. 1971. An exploration in the theory of optimum income taxation. *Review of Economic Studies* **38** 175–208. → pages 61

Noar, P. 1969. The regulation of queue size by levying tolls. *Econometrica* **37** 15–24. → pages 84, 85

Olivella, P. 2003. Shifting public health sector waiting lists to the private sector. *European Journal of Political Economy* **19**(1) 103–132. → pages 3

Pesudovs, K., D. Elliott. 2001. The evolution of cataract surgery. *Optometry Today* 30–32. → pages 42

Prémont, M. 2007. Wait-time guarantees for health services: An analysis of Qubec's reaction to the Chaoulli supreme court decision. *Health Law Journal* **15** 43–86. → pages 1

Quesnel-Vallee, A., M. Bourque, C. Fedick, A. Maioni. 2006. In the aftermath of Chaoulli vs. Quebec: Whose opinion prevailed? *Canadian Medical Association Journal* **175**(9) 1051. → pages 2, 7

Rickman, N., A. McGuire. 1999. Regulating providers' reimbursement in a mixed market for health care. *Scottish Journal of Political Economy* **46**(1) 53–71. → pages 32

Roberts, K. 1971. The theoretical limits to redistribution. *Review of Economic Studies* **51** 177–195. → pages 61

Sanmartin, C., S.E. Shortt, M.L. Barer, S. Sheps, S. Lewis, P.W. McDonald. 2000. Waiting for medical services in Canada: Lots of heat, but little light. *Canadian Medical Association Journal* **162**(9) 1305–1310. → pages 1, 2

Shimkin, N., A. Mandelbaum. 2004. Rational abandonment from tele-queues: Nonlinear waiting cost with heterogeneous preferences. *Queueing Systems* **47** 117–146. → pages 85

Siciliani, L., J. Hurst. 2005. Tackling excessive waiting times for elective surgery: A comparative analysis of policies in 12 OECD countries. *Health Policy* **72**(2) 201–215. → pages 2, 7, 9, 16, 17, 83, 100

Siciliani, L., S. Martin. 2007. An empirical analysis of the impact of choice on waiting times. *Health Economics* **16**(8) 763–779. → pages 24

Siciliani, L., R. Verzulli. 2009. Waiting times and socioeconomic status among elderly Europeans: Evidence from SHARE. *Health Economics* **18**(11) 1295–1306. → pages 100

Tuohy, C.H., C.M. Flood, M. Stabile. 2004. How does private finance affect public health care systems? Marshalling the evidence from OECD nations. *Journal of Health Politics, Policy and Law* **29**(3) 359–396. → pages 2, 9

Veeraraghavan, S., L. Debo. 2009. Joining longer queues: Information externalities in queue choice. *Manufacturing & Service Operations Management* **11**(4) 543–562. → pages 85

Willcox, S., M. Seddon, S. Dunn, R.T. Edwards, J. Pearse, J.V. Tu. 2007. Measuring and reducing waiting times: A cross-national comparison of strategies. *Health Affairs (Project Hope)* **26**(4) 1078–1087. → pages 2, 100

Williams, T. 1985. Nonpreemptive multi-server priority queues. *The Journal of the Operational Research Society* **31**(12) 1105–1107. → pages 103

Wonderling, D., R. Gruen, N. Black. 2005. *Introduction to Health Economics*. Open University Press, Berkshire, England. → pages 60

Wooldridge, J.M. 2009. *Introductory Econometrics: A Modern Approach.* South-Western Cengage Learning, Mason, Ohio. → pages 21, 24

Zweifel, P., F. Breyer, M. Kifmann. 2009. *Health Economics*. 2nd ed. Springer-Verlag, Heidelberg, Germany. → pages 60

Appendix A

Appendix for Chapter 2

A.1 Variable Selection and Calculation

Table A.1 below shows the number of surgeries included in our data versus the total number of surgeries performed in the same period (admission date from April 1, 2005 to March 31, 2007).

The dependent variable - patients' mean waiting time - can be constructed in two manners: prospective or retrospective. The prospective mean waiting time takes the mean of the waiting times of patients incoming in a particular calendar month, while the retrospective mean waiting time takes the mean of the waiting times of patients departing in a particular month. In a stationary queuing system, the long run average waiting time is the same whether it is measured in a prospective or retrospective manner. However, when we construct the observations on a monthly basis, the difference does matter. We decide to use the prospective mean waiting time due to reasons as follows. First, the lagging effects of arrival rates can be captured by prospective waiting time. In contrast, we can construct examples to show that retrospective waiting time is negatively correlated to lags of arrival rate. Second, when we consider patient's choice in the demand function, it is more reasonable to assume that patients make their decisions based on the waiting times that they expect to experience rather than the waiting times in the past. In accordance with prospective waiting time, all monthly statistics are calculated based on decision date.

We recognize the data truncation problem when we measure the monthly arrival rates and the mean waiting times in a prospective manner. To proceed with our study,

Table A.1: Number of surgeries in our data set versus total number of hospitalizations[a]

Hip replacement	2005/6			2006/7		
Province	CJRR*	Total	%	CJRR*	Total	%
Alberta	966	2,846	33.9%	991	2,649	37.4%
British Columbia	1,303	4,237	30.8%	1,701	4,656	36.5%
Manitoba	251	1,248	20.1%	487	1,302	37.4%
New Brunswick	355	656	54.1%	388	651	59.6%
Newfoundland	130	372	34.9%	138	336	41.1%
Nova Scotia	218	935	23.3%	200	831	24.1%
Ontario	435**	12,103	3.6%	1,043	12,494	8.3%
Quebec	681	4,411	15.4%	795	N/A	N/A
Saskatchewan	422	1,059	39.8%	424	1,131	37.5%
Knee replacement	2005/6			2006/7		
Province	CJRR*	Total	%	CJRR*	Total	%
Alberta	1,484	4,001	37.1%	1,499	4,003	37.4%
British Columbia	1,328	5,374	24.7%	1,870	6,446	29.0%
Manitoba	572	1,879	30.4%	1,044	2,202	47.4%
New Brunswick	642	1,284	50.0%	681	968	70.4%
Newfoundland	172	491	35.0%	200	518	38.6%
Nova Scotia	338	1,041	32.5%	349	1,126	31.0%
Ontario	688**	18,990	3.6%	1,642	20,742	7.9%
Quebec	939	5,865	16.0%	1,100	N/A	N/A
Saskatchewan	802	1,497	53.6%	774	1,620	47.8%

[a] Source: Hospital Discharge Database and CJRR. *: CJRR records that do not contain waiting time information are not released to us. **: number of surgeries from October 1, 2005 to March 31, 2006.

we assume data integrity for a certain number of months for each cross section unit, i.e., we try to construct time series data for each cross section unit with certain level of comparability. The way we determine the number of valid months for each cross section unit follows: we pool all waiting time data of a cross section unit and look at the cumulative percentages by the number of months, and the number of months with 90% completion is set as the cut-off point. For instance, there are a total of

1,960 records for cross section unit “Alberta-hip”, 90% of which have a waiting time less than or equal to 11 months, and thus we can assume that for patients arriving in the first 13 months, at least 90% of them are served by the end of the 24-month period. Therefore, the first 13 months of data, i.e., April 2005 to April 2006, will be used as time series data for cross section unit “Alberta-hip”. We can do the similar manipulations for other cross section units. Table A.2 shows the number of months for each cross section unit.

Table A.2: Number of valid months

Province	Hip	Knee
Alberta	13	9
British Columbia	10	7
Manitoba	7	4
New Brunswick	13	11
Newfoundland	18	18
Nova Scotia	10	8
Ontario	11	11
Quebec	15	14
Saskatchewan	6	0

We also need to measure the arrival rates prior to April 2005 to be used as lag variables. For instance, we need to estimate the arrival rate of March 2005. In the data set, patient records with decision date of March 2005 are subject to truncation. To reasonably approximate the arrival rate of March 2005, we use the cumulative percentages of waiting times again. For instance, for “Alberta-hip”, there are 55 patient records with decision date of March 2005 in our data. The waiting time analysis shows that for “Alberta-hip”, 14% of patients are served in less than or equal to one month, then these 55 patients can be seen as the remaining 86% of the cohort, so the approximated number of patient records with decision date of March 2005 is 55/(100%-14%) = 64.

Previous empirical studies also suggest the following explanatory variables to the public waiting times:

(i) Admission threshold (rationing): a surgeon’s admission threshold is set such that patients with severities below that threshold are not added to the waiting list. Less patients will be added to the waiting lists if the threshold is set at a higher level.

(ii) Funding schemes: it is found that the way in which hospitals and physicians are funded also affects a health system's performance. Hospital funding schemes (e.g., fixed, activity-based and performance-based) and the physician payment schemes (Siciliani and Hurst, 2005) are often considered in empirical studies. Siciliani and Hurst (2005) show that fee-for-service payment schemes give physicians more incentives to increase their productivities and discourage the formation of visible queues because of competitive pressure and the incentive to disguise demand, especially when there are no gatekeepers and surgeons assume primary care responsibilities for patients. Siciliani and Hurst (2005) also show that productivity depends, among other things, on the way in which physicians and hospitals are paid. In Canada, all physicians are paid on a fee-for-service basis. We do not have the information about hospital funding schemes.

(iii) Patient's education level: Siciliani and Verzulli (2009) find that higher education levels are associated with shorter waiting times. The authors hypothesize that patients with higher levels of education access more and better health care by having better knowledge, information and skills of complaint.

(iv) Surgeon's prioritization of waiting list: patient prioritization is suggested in Siciliani and Hurst (2005) and Willcox et al. (2007). Patients with different health conditions and different levels of emergency require different amounts of treatment and diagnosis, which would affect a surgeon's prioritization decision. For instance, it has been shown that knee replacement patients with BMI=40 have higher revision rates after a minimum follow-up period of five years. For the sake of cost containment, a surgeon may purposely delay the surgeries for patients with higher levels of obesity so that these patients do not need revisions after the primary replacement.

Due to data limitations, we are unable to include the above explanatory variables in the empirical analysis. Additionally, Table A.3 and A.4 below show the ratios of general physicians to orthopedic surgeons.

Table A.3: Ratios of general physicians to orthopedic surgeons by province

Province	Number of eligible orthopedic surgeons	Number of family medicine doctors		The ratio of family medicine doctors to orthopedic surgeons		Mean waiting time (in days), 2005-2007	
		Fee-for-service count	Full-time equivalent	Fee-for-service count	Full-time equivalent	Hip surgery	Knee surgery
Alberta	54	3,363	2,643	62	49	155	233
British Columbia	97	4,966	3,638	51	38	212	261
Manitoba	24	953	757	40	32	269	314
New Brunswick	27	713	471	26	17	176	216
Newfoundland	15	595	384	40	26	95	109
Nova Scotia	27	950	636	35	24	236	279
Ontario	241	11,301	8,808	47	37	106	113
Quebec	193	7,341	5,513	38	29	143	169
Saskatchewan	24	1,010	750	42	31	339	455

Note: (1) the numbers of eligible orthopedic surgeons are estimated by CJRR for the year of 2005-2006; (2) the source of numbers of family medicine doctors is National Physician Database, 2005-2006 Data Release, Canadian Institute for Health Information; (3) the number of family medicine doctors by fee-for-service count is the number of physicians that registered to receive fee-for-service payments; (4) the number of family medicine doctors by full-time equivalent estimates the number of physicians working in a full-time capacity, which is a weighted count based on the yearly amount of fee-for-service payments received.

Table A.4: Correlation coefficients

		Ratio of general physicians to orthopedic surgeons	
		Fee-for-service count	Full-time equivalent
Mean waiting time	Hip surgery	- 0.120	- 0.069
	Knee surgery	- 0.003	0.040

Appendix B

Appendix for Chapter 3

B.1 Waiting Time of $M/G/n$ Queue with Non-preemptive Priority and Linear Waiting Time Approximation

According to Williams (1985), the average waiting times of $M/G/n$ queue with non-preemptive priority and two classes of customers can be written as follows. We assume that the service times of both classes of customers follow the same distribution. The justification of this assumption is stated in Section 3.4. We define:

$$\begin{aligned}
\lambda_i &= \text{the arrival rate of customers of priority i (i=1,2)} \\
S &= \text{the random variable of service time} \\
n &= \text{number of servers} \\
\rho_i &= \lambda_i E(S)/n \text{ (i=1,2)} \\
\rho &= \rho_1 + \rho_2 \text{ (assumed to be less than 1)}
\end{aligned}$$

Let π be such that

$$\pi = \left\{ 1 + (1-\rho) \sum_{k=0}^{n-1} \rho^{k-n} \frac{n^{k+1-n}(n-1)!}{k!} \right\}^{-1}$$

The expected waiting time of the priority customers ($i = 1$) is

$$w_1 = \frac{E(S^2)\pi}{2E(S)n(1-\rho_1)}$$

It is straightforward to see that w_1 is strictly increasing in λ_1, λ_2 and $E(S)$. The expected waiting time of the non-priority customers ($i = 2$) is

$$w_2 = \frac{w_1}{(1-\rho)}$$

Again, w_2 is strictly increasing in λ_1, λ_2 and $E(S)$. However, we have $\frac{\partial w_2}{\partial \lambda_1} > \frac{\partial w_1}{\partial \lambda_1}$ and $\frac{\partial w_2}{\partial \lambda_2} > \frac{\partial w_1}{\partial \lambda_2}$. For instance, when $n = 1$, we have

$$w_1 = \frac{(\lambda_1+\lambda_2)E(S^2)}{2(1-\rho_1)} \quad w_2 = \frac{(\lambda_1+\lambda_2)E(S^2)}{2(1-\rho_1)(1-\rho_1-\rho_2)}$$

The Taylor series for w_1 and w_2 are as follows:

$$w_1 = \frac{(\lambda_1+\lambda_2)E(S^2)}{2}\left\{1+\rho_1+\rho_1^2+\ldots\right\} = \frac{E(S^2)}{2}\left\{\lambda_1+\lambda_2+\lambda_1(\lambda_1+\lambda_2)E(S)+\ldots\right\}$$

$$w_2 = \frac{(\lambda_1+\lambda_2)E(S^2)}{2}\left\{1+\rho_1+\rho_1^2+\ldots\right\}\left\{1+(\rho_1+\rho_2)+(\rho_1+\rho_2)^2+\ldots\right\}$$

Therefore, w_1 can be approximated by linear terms of arrival rates as:

$$w_1 \sim \frac{E(S^2)}{2}(\lambda_1+\lambda_2)$$

Accordingly, w_2 can be written as:

$$w_2 = w_1\left\{1+(\rho_1+\rho_2)+(\rho_1+\rho_2)^2+\ldots\right\}$$

We know that $\left\{1+(\rho_1+\rho_2)+(\rho_1+\rho_2)^2+\ldots\right\}$ is always greater than 1. Let $\alpha = \frac{E(S^2)}{2}$ and $\beta = \alpha\left[1+(\rho_1+\rho_2)+(\rho_1+\rho_2)^2+\ldots\right]$, then we have linear approximations for w_1 and w_1 as:

$$w_1 \sim \alpha(\lambda_1 + \lambda_2) \quad w_2 \sim \beta(\lambda_1 + \lambda_2)$$

where $\beta > \alpha$.

The expected waiting time of an $M/G/1$ queue (assuming that the service time follows the same distribution as S) under FIFO service principle is:

$$w = \frac{\lambda E(S^2)}{2(1 - \lambda E(S))}$$

Similarly, the linear approximation for w can be written:

$$w \sim \alpha\lambda$$

It should be noted that these linear approximations are applicable only when the traffic intensities ρ_1, ρ_2 and ρ are not close to 1.

B.2 Development of Equilibrium Arrival Rates

The indifferent patient type between queue A and B is:

$$R_A - h \cdot w_A \begin{pmatrix} > \\ = \\ < \end{pmatrix} R_B - h \cdot w_B \quad \Rightarrow \quad h \begin{pmatrix} > \\ = \\ < \end{pmatrix} \frac{R_A - R_B}{w_A - w_B} \tag{B.1}$$

The indifferent patient type between queue A and C is:

$$R_A - h \cdot w_A \begin{pmatrix} > \\ = \\ < \end{pmatrix} R_C - h \cdot w_C \quad \Rightarrow \quad h \begin{pmatrix} > \\ = \\ < \end{pmatrix} \frac{R_A - R_C}{w_A - w_C} \tag{B.2}$$

The indifferent patient between queue B and queue C is:

$$R_B - h \cdot w_B \begin{pmatrix} > \\ = \\ < \end{pmatrix} R_C - h \cdot w_C \quad \Rightarrow \quad h \begin{pmatrix} > \\ = \\ < \end{pmatrix} \frac{R_B - R_C}{w_B - w_C} \tag{B.3}$$

Last, there are indifferent patient types between joining a queue and seeking an outside service with zero reservation net benefit.

$$R_A - h \cdot w_A \begin{pmatrix} > \\ = \\ < \end{pmatrix} 0 \quad \Rightarrow \quad h \begin{pmatrix} > \\ = \\ < \end{pmatrix} \frac{R_A}{w_A} \tag{B.4}$$

$$R_B - h \cdot w_B \begin{pmatrix} > \\ = \\ < \end{pmatrix} 0 \quad \Rightarrow \quad h \begin{pmatrix} > \\ = \\ < \end{pmatrix} \frac{R_B}{w_B} \tag{B.5}$$

$$R_C - h \cdot w_C \begin{pmatrix} > \\ = \\ < \end{pmatrix} 0 \quad \Rightarrow \quad h \begin{pmatrix} > \\ = \\ < \end{pmatrix} \frac{R_C}{w_C} \tag{B.6}$$

According to (B.1)–(B.6), in equilibrium patients' choices of queues are as follows:

$$\begin{cases} \text{Join queue A,} & h \in \left[0, \min\left\{\frac{R_A - R_B}{w_A - w_B}, \frac{R_A - R_C}{w_A - w_C}, \frac{R_A}{w_A}\right\}\right]; \\ \text{Join queue B,} & h \in \left[\frac{R_A - R_B}{w_A - w_B}, \min\left\{\frac{R_B - R_C}{w_B - w_C}, \frac{R_B}{w_B}\right\}\right]; \\ \text{Join queue C,} & h \in \left[\max\left\{\frac{R_A - R_C}{w_A - w_C}, \frac{R_B - R_C}{w_B - w_C}\right\}, \frac{R_C}{w_C}\right]; \\ \text{Choose outside service,} & h \in \left[\max\left\{\frac{R_A}{w_A}, \frac{R_B}{w_B}, \frac{R_C}{w_C}\right\}, +\infty\right). \end{cases} \tag{B.7}$$

(B.7) can be further simplified by the following two lemmas:

Lemma B.1 *In an equilibrium (if exists) where arrival rates to all three queues are positive, we must have*

(i) $\frac{R_A - R_C}{w_A - w_C} \leq \frac{R_A}{w_A}$;

(ii) $\frac{R_B - R_C}{w_B - w_C} \leq \frac{R_B}{w_B}$;

(iii) $\max\left\{\frac{R_B - R_C}{w_B - w_C}, \frac{R_A - R_C}{w_A - w_C}\right\} \leq \frac{R_C}{w_C}$.

Proof: For Case (i), suppose not and we have $\frac{R_A - R_C}{w_A - w_C} > \frac{R_A}{w_A}$ in equilibrium. For patients with $h \in \left[\frac{R_A}{w_A}, \frac{R_A - R_C}{w_A - w_C}\right]$, by (B.1) and (B.4), we have $R_C - hw_C < R_A - hw_A < 0$, so these patients prefer outside service to joining queue C. For patients with $h \in \left(\frac{R_A-R_C}{w_A-w_C}, +\infty\right)$, we have $R_C - hw_C < 0$, i.e. these patients prefer outside service to joining queue C. Therefore, no patients join queue C, which contradicts the assumption that $\lambda_C > 0$. By the same rationale, Case (ii) and Case (iii) follow. ■

Lemma B.2 *In an equilibrium (if exists) where arrival rates to all three queues are positive, we must have:*

$$\frac{R_A - R_B}{w_A - w_B} < \frac{R_A - R_C}{w_A - w_C} < \frac{R_B - R_C}{w_B - w_C}.$$

Proof: Suppose not and we have $\frac{R_A-R_B}{w_A-w_B} \geq \frac{R_B-R_C}{w_B-w_C}$ in equilibrium, then by (B.7) we have $\lambda_B = 0$, which contradicts the assumption that all arrival rates are positive. Also, we have the following derivations:

$$\begin{aligned}
\frac{R_A - R_B}{w_A - w_B} < \frac{R_B - R_C}{w_B - w_C} \Rightarrow \quad & R_A(w_B - w_C) + R_C(w_A - w_B) < R_B(w_A - w_C) \\
\Rightarrow \quad & R_A(w_A - w_C - w_A + w_B) < R_B(w_A - w_C) - R_C(w_A - w_B) \\
\Rightarrow \quad & R_A(w_A - w_C) - R_B(w_A - w_C) < R_A(w_A - w_B) - R_C(w_A - w_B) \\
\Rightarrow \quad & \frac{R_A - R_B}{w_A - w_B} < \frac{R_A - R_C}{w_A - w_C}
\end{aligned}$$

$$\begin{aligned}
\frac{R_A - R_B}{w_A - w_B} < \frac{R_B - R_C}{w_B - w_C} \Rightarrow \quad & R_A(w_B - w_C) + R_C(w_A - w_B) < R_B(w_A - w_C) \\
\Rightarrow \quad & R_A(w_B - w_C) - R_C(w_B - w_C - w_A + w_C) < R_B(w_A - w_C) \\
\Rightarrow \quad & R_A(w_B - w_C) - R_C(w_B - w_C) < R_B(w_A - w_C) - R_C(w_A - w_C) \\
\Rightarrow \quad & \frac{R_A - R_C}{w_A - w_C} < \frac{R_B - R_C}{w_B - w_C}
\end{aligned}$$

■

By Lemma B.1 and B.2, (B.7) can be simplified as follows:

$$\begin{cases} \text{Join queue A,} & h \in \left[0, \frac{R_A-R_B}{w_A-w_B}\right]; \\ \text{Join queue B,} & h \in \left[\frac{R_A-R_B}{w_A-w_B}, \frac{R_B-R_C}{w_B-w_C}\right]; \\ \text{Join queue C,} & h \in \left[\frac{R_B-R_C}{w_B-w_C}, \frac{R_C}{w_C}\right]; \\ \text{Choose outside service,} & h \in \left[\frac{R_C}{w_C}, +\infty\right). \end{cases}$$

and the corresponding arrival rates are as follows:

$$\begin{cases} \lambda_A = F\left(\frac{R_A-R_B}{w_A-w_B}\right)\Lambda \\ \lambda_B = \left\{F\left(\frac{R_B-R_C}{w_B-w_C}\right) - F\left(\frac{R_A-R_B}{w_A-w_B}\right)\right\}\Lambda \\ \lambda_C = \left\{F\left(\frac{R_C}{w_C}\right) - F\left(\frac{R_B-R_C}{w_B-w_C}\right)\right\}\Lambda \end{cases} \tag{B.8}$$

B.3 Proofs

B.3.1 Proof of Proposition 3.1

Proof: First, we assert that for an equilibrium in which none of the queues is empty, we must have $w_A \geq w_B \geq w_C$. For instance, suppose not and we have $w_B < w_C$ in equilibrium. For any patient type h, we therefore have $R_B - h \cdot w_B > R_C - h \cdot w_C$, i.e. all patients prefer queue B to queue C. In this case, queue C would be empty, which contradicts the assumption that $\lambda_C > 0$. By the same rationale, $w_A \geq w_B$ and $w_A \geq w_C$ should follow.

However, if we allow that the arrival rate to some queue could be zero in equilibrium, then the conclusion $w_A \geq w_B \geq w_C$ may not hold. For instance, let Q_a, Q_d and p be such that $R_A = Q_a > R_B = Q_d > R_C = Q_d - p$, then $w_C = w_0(\lambda_0, \lambda_1)$, $w_B = w_1(\lambda_0, \lambda_1)$ and $w_A = w_2(\lambda_2)$. We can construct an equilibrium in which $w_B > w_A$ prevails. Let the service times of both types of physicians follow the same distribution, and the waiting times take the form of $M/G/1$ queue with postponable priority (Appendix B.1), i.e. $n_d = n_a = 1$. Therefore, we have $w_A = w_2 = \frac{\lambda_2 E(S^2)}{2(1-\rho_2)}$,

$w_B = w_1 = \frac{(\lambda_0+\lambda_1)E(S^2)}{2(1-\rho_0)(1-\rho_0-\rho_1)} \geq \frac{(\lambda_0+\lambda_1)E(S^2)}{2(1-\rho_0)^2} > w_C = w_0 = \frac{(\lambda_0+\lambda_1)E(S^2)}{2(1-\rho_0)}$. Let patients be homogeneous and $\theta = 1$ be patient's time cost, service qualities be $Q_a = 1$ and $Q_d = 0.9$, and total patient arrival rate be $\Lambda = 1$. We assume that the service times of both physician types follow the same distribution. Let the first and second moments of service time be $E(S) = E(S^2) = 1$. For any $\lambda_0 \in (0.43, 0.475), \lambda_1 = 0, \lambda_2 = 1-\lambda_0, p = \frac{1-\lambda_0}{2\lambda_0} - \frac{\lambda_0}{2(1-\lambda_0)-0.1}$ constitutes an equilibrium. For any of these equilibria, we have $w_1 > w_2 > w_0$, i.e. $w_B > w_A > w_C$.

The second example uses the same assumptions, parameters and waiting time functions as in the preceding paragraph, except that we use $Q_d = 2, Q_a = 1, p = 1.5$ and $\Lambda = 0.5$ in this example. In this case, $R_A = Q_d > R_B = Q_a > R_C = Q_d - p$ and $w_A = w_1(\lambda_0, \lambda_1)$, $w_B = w_2(\lambda_2)$ and $w_C = w_0(\lambda_0, \lambda_1)$. The resulted equilibrium is that all patients are served in queue 1 ($\lambda_1 = \Lambda = 0.5$), while queue 0 and queue 2 are empty ($\lambda_0 = \lambda_2 = 0$). The waiting times are $w_A = w_1 = w_C = w_0 = 0.5 > w_B = 0$. ■

B.3.2 Proof of Proposition 3.2

Proof: We only prove for the case $Q_d > Q_a > Q_d - p > 0$. Proofs of the other two cases follow the same rationale. We assume that the total arrival rate is Λ and patient's time cost $h \in \left[0, \overline{h}\right]$, where $\overline{h}$ can be $+\infty$. We assume that the outside service has a reservation benefit of zero.

Before going into the details of the proof, we have two important conclusions. First, we assert that $\lambda_1 > 0$. Given any $w_1 \geq 0$, patients with time costs $h \in \left[0, \frac{Q_d}{w_1}\right]$ would join queue 1, so the arrival rate to queue 1 is always positive. Second, we assert that if $\lambda_0 > 0$, then $\lambda_2 > 0$. Suppose not, then we must have $w_0 > 0$ and $w_2 = 0$. In that case, any patient would have a higher net benefit by joining queue 2 than by joining queue 0. Therefore, all these class 0 patients would join queue 2 instead of queue 0, which leads to $\lambda_0 = 0$. This is a contraction to the assumption $\lambda_0 > 0$.

There could exist different types of equilibria. The first type of equilibrium is that only one queue has positive arrival rate, while other two queues are empty. According to the p, the queue with positive arrival rate must be queue 1. Queue 1 accommodate all patients and no patients want to switch to another queue or outside service. The necessary and sufficient conditions for this equilibrium to occur is:

$$Q_d - \overline{h} \cdot w_1(0, \Lambda) \geq Q_a \tag{B.9}$$

The second type of equilibrium is that the arrival rates of two queues are positive and the last queue is empty. According to the above conclusions, the empty queue must be queue 0, i.e. $\lambda_0 = 0$. However, since the arrival rate to queue 1 is positive, i.e. $\lambda_1 > 0$, the virtual waiting time of queue 0 would be positive, $w_0 > 0$. The necessary and sufficient conditions for this scenario to occur is:

$$\begin{cases} \lambda_1 = F\left(\frac{Q_d - Q_a}{w_1 - w_2}\right)\Lambda; \\ \lambda_2 = \left[F\left(\frac{Q_a}{w_2}\right) - F\left(\frac{Q_d - Q_a}{w_1 - w_2}\right)\right]\Lambda; \\ Q_d - p - \frac{Q_a}{w_2} w_0(0, \lambda_1) \leq 0. \end{cases} \tag{B.10}$$

Given that neither condition (B.9) nor condition (B.10) is satisfied, the arrival rates to all three queues are positive, i.e. none of the queue is empty. Our task below is to prove that there exist non-negative arrival rates $\lambda_0 > 0$, $\lambda_1 > 0$ and $\lambda_2 > 0$ that solve the following simultaneous equations:

$$\lambda_1 = F\left(\frac{Q_d - Q_a}{w_1 - w_2}\right)\Lambda, \tag{B.11}$$

$$\lambda_2 = \left[F\left(\frac{Q_a - Q_d + p}{w_2 - w_0}\right) - F\left(\frac{Q_d - Q_a}{w_1 - w_2}\right)\right]\Lambda, \tag{B.12}$$

$$\lambda_0 = \left[F\left(\frac{Q_d - p}{w_0}\right) - F\left(\frac{Q_a - Q_d + p}{w_2 - w_0}\right)\right]\Lambda. \tag{B.13}$$

The idea of the proof is as follows. First, we show that there exist nonnegative triplets $(\lambda_0, \lambda_1, \lambda_2)$, denoted as set **G1**, that solve equation (B.11). Next, we show that there exists a subset of **G1**, denoted as **G2**, that solve equation (B.12). Finally, we show that there exists $(\lambda_0, \lambda_1, \lambda_2)$ in **G2** that solve equation (B.12).

First, we show that **G1** is not empty. We look at triplets $(0, \lambda_1, 0)$. In this case, we have $w_1 = w_1(0, \lambda_1)$ and $w_2 = 0$. Equation (B.11) then becomes $\lambda_1 = F\left(\frac{Q_d - Q_a}{w_1}\right)\Lambda$. If $\lambda_1 = 0$, we have $\lambda_1 < F\left(\frac{Q_d - Q_a}{w_1}\right)\Lambda = \Lambda$; if $\lambda = \Lambda$, we have $\Lambda \geq F\left(\frac{Q_d - Q_a}{w_1}\right)\Lambda$. By continuity of $\lambda_1 \in [0, \Lambda]$, there must exist $\lambda_1 > 0$ that solve $\lambda_1 = F\left(\frac{Q_d - Q_a}{w_1}\right)\Lambda$. Therefore, **G1** is not empty. For each given λ_1 in **G1**, there could exist multiple nonnegative (λ_0, λ_2) that solve (B.11).

Now let us discuss the range of λ_1 in **G1**. The lower bound of λ_1 in **G1**, denoted as $\underline{\lambda}_1$, solves $\lambda_1 = F\left(\frac{Q_d - Q_a}{w_1}\right)\Lambda$, where $w_1 = w_1(\Lambda - \lambda_1, \lambda_1)$. The upper bound of λ_1 in **G1**, denoted as $\overline{\lambda}_1$, solves $\lambda_1 = F\left(\frac{Q_d - Q_a}{w_1 - w_2}\right)\Lambda$, where $w_1 = w_1(0, \lambda_1)$ and $w_2 = w_2(\Lambda - \lambda_1)$.

Next, we show that there exist a subset of **G1** that solve (B.12). For every triplet $(\lambda_0, \lambda_1, \lambda_2)$ in **G1**, (B.12) can be simplified as

$$\lambda_2 + \lambda_1 = F\left(\frac{Q_a - Q_d + p}{w_2 - w_0}\right)\Lambda \tag{B.14}$$

First, by substituting $\overline{\lambda}_1$ and $\lambda_2 = \Lambda - \overline{\lambda}_1$ into both sides of (B.14), we have $\Lambda \geq F\left(\frac{Q_a - Q_d + p}{w_2 - w_0}\right)\Lambda$, where $w_2 = w_2(\Lambda - \overline{\lambda}_1)$ and $w_0 = w_0(0, \overline{\lambda}_1)$. Next, we assert that there exist triplet $(\lambda_0, \lambda_1, \lambda_2)$ in **G1** so as to make $w_2 < w_0$. For instance, $(\underline{\lambda}_1, \Lambda - \underline{\lambda}_1, 0)$ makes $w_2 = 0 < w_0 = w_0(\Lambda - \underline{\lambda}_1, \underline{\lambda}_1)$. However, if all $(\lambda_0, \lambda_1, \lambda_2)$ in **G1** make $w_2 \leq w_0$, then it must be that the given parameters satisfy condition (B.10), which contradicts our assumption. Therefore, there exist $(\lambda_0, \lambda_1, \lambda_2)$ in **G1** such that $w_2 \geq w_0$. By continuity of $(\lambda_0, \lambda_1, \lambda_2)$, there must exist $(\lambda_0, \lambda_1, \lambda_2)$ in **G1** such that $w_2 = w_0$. Substituting these $(\lambda_0, \lambda_1, \lambda_2)$ into (B.14), we obtain $\lambda_2 + \lambda_1 < F\left(\frac{Q_a - Q_d + p}{w_2 - w_0}\right)\Lambda = \Lambda$. In conclusion, by continuity of $(\lambda_0, \lambda_1, \lambda_2)$, there must exist $(\lambda_0, \lambda_1, \lambda_2)$ in **G1** that satisfy $\lambda_2 + \lambda_1 = F\left(\frac{Q_a - Q_d + p}{w_2 - w_0}\right)\Lambda$. We denote this set of $(\lambda_0, \lambda_1, \lambda_2)$ as **G2**.

Lastly, we prove that there exist $(\lambda_0, \lambda_1, \lambda_2)$ in **G2** that satisfy (B.13). For every $(\lambda_0, \lambda_1, \lambda_2)$ in **G2**, equation (B.13) can be simplified as:

$$\lambda_0 + \lambda_1 + \lambda_2 = F\left(\frac{Q_d - p}{w_0}\right)\Lambda \tag{B.15}$$

If for every $(\lambda_0, \lambda_1, \lambda_2)$ in **G2**, we always have $\lambda_0 + \lambda_1 + \lambda_2 < F\left(\frac{Q_d - p}{w_0}\right)\Lambda$, then it means that we can always serve more patients than $\lambda_0 + \lambda_1 + \lambda_2$. Keeping doing that would make the total number of patients being served in the health system equal to Λ. Therefore, we can conclude that there must exist triplet $(\lambda_0, \lambda_1, \lambda_2)$ in **G2** that satisfy $\lambda_0 + \lambda_1 + \lambda_2 \geq F\left(\frac{Q_d - p}{w_0}\right)\Lambda$. On the other hand, we can find triplets $(0, \lambda_1, \lambda_2)$ that solve:

$$\begin{cases} \lambda_1 = F\left(\frac{Q_d - Q_a}{w_1 - w_2}\right)\Lambda; \\ \lambda_2 = \left[F\left(\frac{Q_a - Q_d + p}{w_2}\right) - F\left(\frac{Q_d - Q_a}{w_1 - w_2}\right)\right]\Lambda. \end{cases} \tag{B.16}$$

Note that the solutions $(0, \lambda_1, \lambda_2)$ above belongs to **G2**. Since we assume that condition (B.10) is not satisfied, we must have $Q_d - p - \frac{Q_a - Q_d + p}{w_2} \cdot w_0(0, \lambda_1) > 0$.

That means, there exist $(\lambda_0,\lambda_1,\lambda_2)$ in **G2** that satisfy $\lambda_0+\lambda_1+\lambda_2 < F\left(\frac{Q_d-p}{w_0}\right)\Lambda$. By continuity of $(\lambda_0,\lambda_1,\lambda_2)$ in **G2**, there must exist $(\lambda_0,\lambda_1,\lambda_2)$ in **G2** that satisfy $\lambda_0+\lambda_1+\lambda_2 = F\left(\frac{Q_d-p}{w_0}\right)\Lambda$.

In conclusion, we have established the existence of nonnegative arrival rates $(\lambda_0,\lambda_1,\lambda_2)$ that solve simultaneous equations (B.11) – (B.13). Since simultaneous equations (B.11) – (B.13) plus nonnegativeness are the sufficient and necessary conditions of the existence of equilibrium, we have established the existence of equilibrium. ■

B.3.3 Proof of Proposition 3.3

Proof: Similar to the proof of Proposition 3.2, we only prove for the case $Q_d > Q_a > Q_d - p$. Proofs of the other two cases follow the same rationale.

The waiting times take linear forms $w_0 = \alpha(\lambda_0+\lambda_1)$, $w_1 = \beta(\lambda_0+\lambda_1)$ and $w_2 = \alpha\lambda_2$, where $\alpha < \beta$. Suppose that there are two equilibria: $(\lambda_0,\lambda_1,\lambda_2)$ and $\left(\tilde{\lambda}_0,\tilde{\lambda}_1,\tilde{\lambda}_2\right)$. Therefore, $(\lambda_0,\lambda_1,\lambda_2)$ and $\left(\tilde{\lambda}_0,\tilde{\lambda}_1,\tilde{\lambda}_2\right)$ must satisfy the following two sets of simultaneous equations:

$$\begin{cases} \lambda_1 = F\left(\frac{Q_d-Q_a}{\beta(\lambda_0+\lambda_1)-\alpha\lambda_2}\right)\Lambda; \\ \lambda_2 = \left[F\left(\frac{Q_a-Q_d+p}{\alpha(\lambda_2-\lambda_0-\lambda_1)}\right) - F\left(\frac{Q_d-Q_a}{\beta(\lambda_0+\lambda_1)-\alpha\lambda_2}\right)\right]\Lambda; \\ \lambda_0 = \left[F\left(\frac{Q_d-p}{\alpha(\lambda_0+\lambda_1)}\right) - F\left(\frac{Q_a-Q_d+p}{\alpha(\lambda_2-\lambda_0-\lambda_1)}\right)\right]\Lambda. \end{cases}$$

$$\begin{cases} \tilde{\lambda}_1 = F\left(\frac{Q_d-Q_a}{\beta(\tilde{\lambda}_0+\tilde{\lambda}_1)-\alpha\tilde{\lambda}_2}\right)\Lambda; \\ \tilde{\lambda}_2 = \left[F\left(\frac{Q_a-Q_d+p}{\alpha(\tilde{\lambda}_2-\tilde{\lambda}_0-\tilde{\lambda}_1)}\right) - F\left(\frac{Q_d-Q_a}{\beta(\tilde{\lambda}_0+\tilde{\lambda}_1)-\alpha\tilde{\lambda}_2}\right)\right]\Lambda; \\ \tilde{\lambda}_0 = \left[F\left(\frac{Q_d-p}{\alpha(\tilde{\lambda}_0+\tilde{\lambda}_1)}\right) - F\left(\frac{Q_a-Q_d+p}{\alpha(\tilde{\lambda}_2-\tilde{\lambda}_0-\tilde{\lambda}_1)}\right)\right]\Lambda. \end{cases}$$

Without loss of generality, we assume $\lambda_2 < \tilde{\lambda}_2$. If at the same time we also have $\lambda_1 < \tilde{\lambda}_1$, then it must be that $\beta(\lambda_0+\lambda_1)-\alpha\lambda_2 > \beta(\tilde{\lambda}_0+\tilde{\lambda}_1)-\alpha\tilde{\lambda}_2 \Rightarrow \alpha(\tilde{\lambda}_2-\lambda_2) > \beta(\tilde{\lambda}_0+\tilde{\lambda}_1-\lambda_0-\lambda_1)$.

If $(\tilde{\lambda}_0+\tilde{\lambda}_1-\lambda_0-\lambda_1) > 0$, then $\alpha(\tilde{\lambda}_2-\lambda_2) > \beta(\tilde{\lambda}_0+\tilde{\lambda}_1-\lambda_0-\lambda_1) \Rightarrow \alpha(\tilde{\lambda}_2-\lambda_2) > \alpha(\tilde{\lambda}_0+\tilde{\lambda}_1-\lambda_0-\lambda_1) \Rightarrow \alpha(\tilde{\lambda}_2-\tilde{\lambda}_0-\tilde{\lambda}_1) > \alpha(\lambda_2-\lambda_0-\lambda_1)$; otherwise, if $(\tilde{\lambda}_0+\tilde{\lambda}_1-\lambda_0-\lambda_1) < 0$, then we also have $\alpha(\tilde{\lambda}_2-\tilde{\lambda}_0-\tilde{\lambda}_1) > \alpha(\lambda_2-\lambda_0-\lambda_1)$. In this case, we have $\tilde{\lambda}_1+\tilde{\lambda}_2 = F\left(\frac{Q_a-Q_d+p}{\alpha(\tilde{\lambda}_2-\tilde{\lambda}_0-\tilde{\lambda}_1)}\right)\Lambda < \lambda_1+\lambda_2 = F\left(\frac{Q_a-Q_d+p}{\alpha(\lambda_2-\lambda_0-\lambda_1)}\right)\Lambda$, which contradicts the assumption that $\tilde{\lambda}_1+\tilde{\lambda}_2 > \lambda_1+\lambda_2$. In conclusion, $\lambda_1 < \tilde{\lambda}_1$ would not exist in the equilibrium.

On the other hand, if we have $\lambda_1 \geq \tilde{\lambda}_1$ in equilibrium, then it must be that $\beta(\lambda_0+\lambda_1)-\alpha\lambda_2 \leq \beta(\tilde{\lambda}_0+\tilde{\lambda}_1)-\alpha\tilde{\lambda}_2 \Rightarrow 0 < \alpha(\tilde{\lambda}_2-\lambda_2) \leq \beta(\tilde{\lambda}_0+\tilde{\lambda}_1-\lambda_0-\lambda_1) \Rightarrow \tilde{\lambda}_0 > \lambda_0$ and $\tilde{\lambda}_0+\tilde{\lambda}_1 > \lambda_0+\lambda_1$. Since $\tilde{\lambda}_0+\tilde{\lambda}_1 > \lambda_0+\lambda_1 \Rightarrow F\left(\frac{Q_d-p}{\alpha(\tilde{\lambda}_0+\tilde{\lambda}_1)}\right)\Lambda < F\left(\frac{Q_d-p}{\alpha(\lambda_0+\lambda_1)}\right)\Lambda \Rightarrow \tilde{\lambda}_0+\tilde{\lambda}_1+\tilde{\lambda}_2 < \lambda_0+\lambda_1+\lambda_2$, it contradicts the assumption that $\tilde{\lambda}_2 > \lambda_2$ and $\tilde{\lambda}_0+\tilde{\lambda}_1 > \lambda_0+\lambda_1$.

In conclusion, there only exists a unique equilibrium when the waiting times take the linear forms. ■

B.3.4 Proof of Proposition 3.6

Proof: When a common queue is split into multiple queues, some efficiency would be lost due to the argument of resource pooling. Provided that the total arrival rate remains the same before and after the queue separation, the expected waiting time in at least one of the resulted queues would be longer than the expected waiting time in the original common queue. In our case, the total patient arrival rate is Λ whether physician dual practice is allowed or not. If physician dual practice is allowed, the common $M/G/(n_d+n_a)$ queue would be split into a $M/G/n_d$ queue with non-preemptive priority and a $M/G/n_a$ queue under FIFO service principle. The expected waiting time in at least one of these two resulted queues would be longer than the expected waiting time in the common $M/G/(n_d+n_a)$ queue (i.e. $\widetilde{w}$). By Proposition 3.1, we know that $w_A \geq w_B \geq w_C$, so we must have $w_A > \widetilde{w}$. If $Q_a < Q_d$, then we have $w_A = w_1 > \widetilde{w}$; if $Q_a > Q_d$, then we have $w_A = w_2 > \widetilde{w}$. However, we cannot determine whether w_B or w_C is longer than $\widetilde{w}$. ■

B.3.5 Proof of Proposition 3.7

Proof: By Proposition 3.4, we know that if $Q_a < Q_d$, then $\frac{\partial w_1}{\partial p} < 0$. Therefore, $\frac{Q_d-\frac{n_d\widetilde{Q}_d+n_aQ_a}{n_d+n_a}}{w_1-\widetilde{w}}$ is increasing in p, i.e. h_1 is increasing in p, so $\frac{\partial h_1}{\partial p} > 0$. Also by Proposition 3.4, we have $\frac{\partial \lambda_1}{\partial p} > 0$: the higher the price, the more patients are served in the public queue of dual practice physicians. Therefore, h_2 is increasing in p too. Combining these two results, we know that when p increases, more patients in the public queue of dual practice physicians would benefit from service quality enhancement that is induced by allowing physician dual practice. ■

B.3.6 Proof of Proposition 3.8

Proof: In the case when physician dual practice is not allowed, patients are not able to distinguish between physicians with service quality $\widetilde{Q}_d$ and physicians with service quality Q_a. Therefore, incoming patients are served under FIFO service principle, and the expected waiting time is $\widetilde{w}$ regardless of $\widetilde{Q}_d$.

For class 1 patients, at the right hand side of (3.6), $(w_1 - \widetilde{w})$ is constant regardless of $\widetilde{Q}_d$, $\left(Q_d - \frac{n_d\widetilde{Q}_d + n_a Q_a}{n_d + n_a}\right)$ is decreasing in $\widetilde{Q}_d$, so the right hand side of (3.6) is decreasing in $\widetilde{Q}_d$, i.e. the number of class 1 patients who benefit from physician dual practice decreases in $\widetilde{Q}_d$.

The proofs for both class 0 patients and class 2 patients are the same, so we only present the proof for class 0 patients. At both sides of (3.7), $\left(Q_d - p - \frac{n_d\widetilde{Q}_d + n_a Q_a}{n_d + n_a}\right)$ is decreasing in $\widetilde{Q}_d$ and $(w_0 - \widetilde{w})$ is constant regardless of $\widetilde{Q}_d$. Dividing both sides by $(w_0 - \widetilde{w})$, we obtain

$$h \left(\begin{array}{c} \leq \\ > \end{array} \right) \left[Q_d - p - \frac{n_d\widetilde{Q}_d + n_a Q_a}{n_d + n_a} \right] / \left[w_0 - \widetilde{w} \right], \text{ if } \left\{ \begin{array}{l} w_0 - \widetilde{w} \geq 0 \\ w_0 - \widetilde{w} < 0 \end{array} \right. \tag{B.17}$$

The right hand side of (B.17) is a threshold of time cost. If $(w_0 - \widetilde{w}) \geq 0$, patients with time costs lower than this threshold benefit from physician dual practice. In this case, the threshold is decreasing in $\widetilde{Q}_d$, so the number of class 0 patients who benefit from physician dual practice decreases in $\widetilde{Q}_d$. On the other hand, if $(w_0 - \widetilde{w}) <$ 0, patients with time costs higher than this threshold benefit from physician dual practice. In this case, the threshold is increasing in $\widetilde{Q}_d$ – again, the number of class 0 patients who benefit from physician dual practice decreases in $\widetilde{Q}_d$. In conclusion, regardless of the sign of $(w_0 - \widetilde{w})$, the number of class 0 patients who benefit from physician dual practice decreases in $\widetilde{Q}_d$. ■

B.4 Comparative Statics of Equilibrium Arrival Rates and Equilibrium Waiting Times with Respect to Price and Service Quality Difference

The equilibrium arrival rates are defined in (3.2), (3.3) and (3.4). By taking derivative with respect to p at both sides of each equation and reshuffling the items, we have

$$\begin{pmatrix} (3.2) \\ \\ (3.3) \\ \\ (3.4) \end{pmatrix} \xrightarrow{\hspace{3cm}} \text{taking derivative w.r.t. p}$$

$$\underbrace{\begin{pmatrix} 1-\frac{\partial g_1}{\partial \lambda_A} & -\frac{\partial g_1}{\partial \lambda_B} & -\frac{\partial g_1}{\partial \lambda_C} \\ \\ -\frac{\partial g_2}{\partial \lambda_A}+\frac{\partial g_1}{\partial \lambda_A} & 1-\frac{\partial g_2}{\partial \lambda_B}+\frac{\partial g_1}{\partial \lambda_B} & -\frac{\partial g_2}{\partial \lambda_C}+\frac{\partial g_1}{\partial \lambda_C} \\ \\ -\frac{\partial g_3}{\partial \lambda_A}+\frac{\partial g_2}{\partial \lambda_A} & -\frac{\partial g_3}{\partial \lambda_B}+\frac{\partial g_2}{\partial \lambda_B} & 1-\frac{\partial g_3}{\partial \lambda_C}+\frac{\partial g_2}{\partial \lambda_C} \end{pmatrix}}_{\mathbf{M}} \cdot \begin{pmatrix} \frac{\partial \lambda_A}{\partial p} \\ \\ \frac{\partial \lambda_B}{\partial p} \\ \\ \frac{\partial \lambda_C}{\partial p} \end{pmatrix} = \begin{pmatrix} \frac{\partial g_1}{\partial p} \\ \\ \frac{\partial g_2}{\partial p} \\ \\ \frac{\partial g_3}{\partial p} \end{pmatrix}$$

where $g_1(\lambda_A, \lambda_B, \lambda_c) = F\left(\frac{R_A - R_B}{w_A - w_B}\right)\Lambda$, $g_2(\lambda_A, \lambda_B, \lambda_c) = F\left(\frac{R_B - R_C}{w_B - w_C}\right)\Lambda$ and $g_3(\lambda_A, \lambda_B, \lambda_c) = F\left(\frac{R_C}{w_C}\right)\Lambda$. Therefore, the comparative statics of equilibrium arrival rates to p are:

$$\begin{pmatrix} \frac{\partial \lambda_A}{\partial p} \\ \\ \frac{\partial \lambda_B}{\partial p} \\ \\ \frac{\partial \lambda_C}{\partial p} \end{pmatrix} = \mathbf{M}^{-1} \cdot \begin{pmatrix} \frac{\partial g_1}{\partial p} \\ \\ \frac{\partial g_2}{\partial p} \\ \\ \frac{\partial g_3}{\partial p} \end{pmatrix}$$

Using Maple v.11, we check the expression of $\mathbf{M}^{-1}$. Each entry of $\mathbf{M}^{-1}$ contains 11 linear and quadratic terms of $\frac{\partial g}{\partial \lambda}$'s so its sign and functional properties are not easy to determine. Therefore, we need to limit our analysis to some specific case by which we can simplify the expression of $\mathbf{M}^{-1}$.

The specific case for our analysis assumes: (I) $\lambda_0 + \lambda_1 + \lambda_2 = \Lambda$, i.e., there is no outside service; (II) h follows a uniform distribution in $[0,1]$; (III) linear forms of waiting time in Appendix B.1. It is noted that the linear forms of waiting time in Appendix B.1 use the assumption that the service times of both types of physician follow the same distribution. Given $\lambda_0 + \lambda_1 + \lambda_2 = \Lambda$, we have $g_3 = \Lambda$ and $g_3 - g_2 = \Lambda - \lambda_A - \lambda_B$, so $\mathbf{M}$ can be simplified into:

$$\underbrace{\begin{pmatrix} 1-\frac{\partial g_1}{\partial \lambda_A} & -\frac{\partial g_1}{\partial \lambda_B} & -\frac{\partial g_1}{\partial \lambda_C} \\ -\frac{\partial g_2}{\partial \lambda_A}+\frac{\partial g_1}{\partial \lambda_A} & 1-\frac{\partial g_2}{\partial \lambda_B}+\frac{\partial g_1}{\partial \lambda_B} & -\frac{\partial g_2}{\partial \lambda_C}+\frac{\partial g_1}{\partial \lambda_C} \\ 1 & 1 & 1 \end{pmatrix}}_{\mathbf{M}} \cdot \begin{pmatrix} \frac{\partial \lambda_A}{\partial p} \\ \frac{\partial \lambda_B}{\partial p} \\ \frac{\partial \lambda_C}{\partial p} \end{pmatrix} = \begin{pmatrix} \frac{\partial g_1}{\partial p} \\ \frac{\partial g_2}{\partial p} \\ 0 \end{pmatrix}$$

We can calculate the inverse matrix of **M** as

$$\mathbf{M}^{-1} = \frac{1}{L}\begin{pmatrix} 1-\frac{\partial g_2}{\partial \lambda_B}+\frac{\partial g_1}{\partial \lambda_B}+\frac{\partial g_2}{\partial \lambda_C}-\frac{\partial g_1}{\partial \lambda_C} & \frac{\partial g_1}{\partial \lambda_B}-\frac{\partial g_1}{\partial \lambda_C} & \frac{\partial g_1}{\partial \lambda_B}\frac{\partial g_2}{\partial \lambda_C}+\frac{\partial g_1}{\partial \lambda_C}-\frac{\partial g_1}{\partial \lambda_C}\frac{\partial g_2}{\partial \lambda_B} \\ \frac{\partial g_2}{\partial \lambda_A}-\frac{\partial g_1}{\partial \lambda_A}-\frac{\partial g_2}{\partial \lambda_C}+\frac{\partial g_1}{\partial \lambda_C} & 1-\frac{\partial g_1}{\partial \lambda_A}+\frac{\partial g_1}{\partial \lambda_C} & \frac{\partial g_2}{\partial \lambda_C}-\frac{\partial g_1}{\partial \lambda_C}-\frac{\partial g_1}{\partial \lambda_A}\frac{\partial g_2}{\partial \lambda_C}+\frac{\partial g_1}{\partial \lambda_C}\frac{\partial g_2}{\partial \lambda_A} \\ -\frac{\partial g_2}{\partial \lambda_A}+\frac{\partial g_1}{\partial \lambda_A}-1+\frac{\partial g_2}{\partial \lambda_B}-\frac{\partial g_1}{\partial \lambda_B} & -1+\frac{\partial g_1}{\partial \lambda_A}-\frac{\partial g_1}{\partial \lambda_B} & 1-\frac{\partial g_2}{\partial \lambda_B}+\frac{\partial g_1}{\partial \lambda_B}-\frac{\partial g_1}{\partial \lambda_A}+\frac{\partial g_1}{\partial \lambda_A}\frac{\partial g_2}{\partial \lambda_B}-\frac{\partial g_1}{\partial \lambda_B}\frac{\partial g_2}{\partial \lambda_A} \end{pmatrix}$$

$$= \frac{1}{L}\begin{pmatrix} 1-\frac{\partial g_2}{\partial \lambda_B}+\frac{\partial g_2}{\partial \lambda_C} & 0 & \frac{\partial g_1}{\partial \lambda_B}\frac{\partial g_2}{\partial \lambda_C}+\frac{\partial g_1}{\partial \lambda_C}-\frac{\partial g_1}{\partial \lambda_C}\frac{\partial g_2}{\partial \lambda_B} \\ \frac{\partial g_2}{\partial \lambda_A}-1-\frac{\partial g_2}{\partial \lambda_C} & 0 & \frac{\partial g_2}{\partial \lambda_C}-\frac{\partial g_1}{\partial \lambda_C}-\frac{\partial g_1}{\partial \lambda_A}\frac{\partial g_2}{\partial \lambda_C}+\frac{\partial g_1}{\partial \lambda_C}\frac{\partial g_2}{\partial \lambda_A} \\ -\frac{\partial g_2}{\partial \lambda_A}+\frac{\partial g_2}{\partial \lambda_B} & 0 & 1-\frac{\partial g_2}{\partial \lambda_B}+\frac{\partial g_1}{\partial \lambda_B}-\frac{\partial g_1}{\partial \lambda_A}+\frac{\partial g_1}{\partial \lambda_A}\frac{\partial g_2}{\partial \lambda_B}-\frac{\partial g_1}{\partial \lambda_B}\frac{\partial g_2}{\partial \lambda_A} \end{pmatrix} + \frac{1}{L}\begin{pmatrix} \frac{\partial g_1}{\partial \lambda_B}-\frac{\partial g_1}{\partial \lambda_C} & \frac{\partial g_1}{\partial \lambda_B}-\frac{\partial g_1}{\partial \lambda_C} & 0 \\ 1-\frac{\partial g_1}{\partial \lambda_A}+\frac{\partial g_1}{\partial \lambda_C} & 1-\frac{\partial g_1}{\partial \lambda_A}+\frac{\partial g_1}{\partial \lambda_C} & 0 \\ -1+\frac{\partial g_1}{\partial \lambda_A}-\frac{\partial g_1}{\partial \lambda_B} & -1+\frac{\partial g_1}{\partial \lambda_A}-\frac{\partial g_1}{\partial \lambda_B} & 0 \end{pmatrix}$$

where $L = \frac{\partial g_1}{\partial \lambda_B}\frac{\partial g_2}{\partial \lambda_C}-\frac{\partial g_1}{\partial \lambda_C}\frac{\partial g_2}{\partial \lambda_B}-\frac{\partial g_1}{\partial \lambda_B}\frac{\partial g_2}{\partial \lambda_A}+\frac{\partial g_1}{\partial \lambda_C}\frac{\partial g_2}{\partial \lambda_A}+1-\frac{\partial g_2}{\partial \lambda_B}+\frac{\partial g_1}{\partial \lambda_B}+\frac{\partial g_2}{\partial \lambda_C}-\frac{\partial g_1}{\partial \lambda_A}+\frac{\partial g_1}{\partial \lambda_A}\frac{\partial g_2}{\partial \lambda_B}-\frac{\partial g_1}{\partial \lambda_A}\frac{\partial g_2}{\partial \lambda_C}$.

Therefore, the marginal equilibrium arrival rates of price can be written as:

$$\begin{pmatrix} \frac{\partial \lambda_A}{\partial p} \\ \\ \frac{\partial \lambda_B}{\partial p} \\ \\ \frac{\partial \lambda_C}{\partial p} \end{pmatrix} = \mathbf{M}^{-1} \cdot \begin{pmatrix} \frac{\partial g_1}{\partial p} \\ \\ \frac{\partial g_2}{\partial p} \\ \\ 0 \end{pmatrix}$$

Similarly, the marginal equilibrium arrival rates of service quality difference can be written as:

$$\begin{pmatrix} \frac{\partial \lambda_A}{\partial Q} \\ \\ \frac{\partial \lambda_B}{\partial Q} \\ \\ \frac{\partial \lambda_C}{\partial Q} \end{pmatrix} = \mathbf{M}^{-1} \cdot \begin{pmatrix} \frac{\partial g_1}{\partial Q} \\ \\ \frac{\partial g_2}{\partial Q} \\ \\ 0 \end{pmatrix}$$

Under assumptions (I), (II) and (III), we have $g_1(\lambda_A, \lambda_B, \lambda_c) = \frac{R_A - R_B}{w_A - w_B}\Lambda$, $g_2(\lambda_A, \lambda_B, \lambda_c) = \frac{R_B - R_C}{w_B - w_C}\Lambda$, $w_0 = \alpha(\lambda_0 + \lambda_1)$, $w_1 = \beta(\lambda_0 + \lambda_1)$ and $w_2 = \alpha\lambda_2$. Therefore, the marginal equilibrium waiting times of price can be written as $\frac{\partial w_0}{\partial p} = \alpha\left(\frac{\partial \lambda_0}{\partial p} + \frac{\partial \lambda_1}{\partial p}\right)$, $\frac{\partial w_1}{\partial p} = \beta\left(\frac{\partial \lambda_0}{\partial p} + \frac{\partial \lambda_1}{\partial p}\right)$ and $\frac{\partial w_2}{\partial p} = \alpha\frac{\partial \lambda_2}{\partial p}$. The marginal equilibrium waiting times of service quality difference can be written out accordingly, which are skipped here for the sake of conciseness.

In the case when $Q_d > Q_a$ **and** $p < Q_d - Q_a$, we have $g_1 = \frac{p}{w_1 - w_0}\Lambda$, $g_2 = \frac{Q_d - p - Q_a}{w_0 - w_2}\Lambda$, $\lambda_A = \lambda_1$, $\lambda_B = \lambda_0$ and $\lambda_C = \lambda_2$. The comparative statics are:

$$\begin{pmatrix} \partial\lambda_0/\partial p \\ \\ \partial\lambda_1/\partial p \\ \\ \partial\lambda_2/\partial p \end{pmatrix} = \frac{\Lambda/L}{w_1 - w_0} \begin{pmatrix} 2K - 1 \\ \\ 1 - 2K \\ \\ 0 \end{pmatrix} + \left[\frac{\Lambda/L}{w_1 - w_0} - \frac{\Lambda/L}{w_0 - w_2}\right] \begin{pmatrix} 1 - H \\ \\ H \\ \\ -1 \end{pmatrix}$$

$$\begin{pmatrix} \partial\lambda_0/\partial Q \\ \\ \partial\lambda_1/\partial Q \\ \\ \partial\lambda_2/\partial Q \end{pmatrix} = \frac{\Lambda/L}{w_0 - w_2} \begin{pmatrix} 1-H \\ \\ H \\ \\ -1 \end{pmatrix} \quad \Rightarrow \quad \begin{pmatrix} \partial\lambda_0/\partial Q > 0 \\ \\ \partial\lambda_1/\partial Q < 0 \\ \\ \partial\lambda_2/\partial Q < 0 \end{pmatrix} \quad \Rightarrow \quad \frac{\partial w_0}{\partial Q} > 0, \frac{\partial w_1}{\partial Q} > 0$$

where $H = \frac{\partial g_1}{\lambda_A} = \frac{\partial g_1}{\lambda_B} = -\frac{p\Lambda(\beta-\alpha)}{(w_1-w_0)^2} < 0, K = \frac{\partial g_2}{\lambda_A} = \frac{\partial g_2}{\lambda_B} = -\frac{\partial g_2}{\lambda_C} = -\frac{(Q_d-p-Q_a)\Lambda\alpha}{(w_0-w_2)^2} < 0,$
$L = 1 - 2K > 0.$

In the case when $Q_d > Q_a$ **and** $p > Q_d - Q_a$, we have $g_1 = \frac{Q_d - Q_a}{w_1 - w_2}\Lambda$, $g_2 = \frac{Q_a - Q_d + p}{w_2 - w_0}\Lambda$, $\lambda_A = \lambda_1$, $\lambda_B = \lambda_2$ and $\lambda_C = \lambda_0$. The comparative statics are:

$$\begin{pmatrix} \partial\lambda_0/\partial p \\ \\ \partial\lambda_1/\partial p \\ \\ \partial\lambda_2/\partial p \end{pmatrix} = \frac{\Lambda/L}{w_2 - w_0} \begin{pmatrix} -1 + H - J \\ \\ J - H \\ \\ 1 \end{pmatrix} \Rightarrow \begin{pmatrix} \partial\lambda_0/\partial p < 0 \\ \\ \partial\lambda_1/\partial p > 0 \\ \\ \partial\lambda_2/\partial p > 0 \end{pmatrix} \Rightarrow \frac{\partial w_0}{\partial p} < 0, \frac{\partial w_1}{\partial p} < 0$$

$$\begin{pmatrix} \partial\lambda_0/\partial Q \\ \\ \partial\lambda_1/\partial Q \\ \\ \partial\lambda_2/\partial Q \end{pmatrix} = \frac{\Lambda/L}{w_1 - w_2} \begin{pmatrix} -2K \\ \\ 1 + 2K \\ \\ -1 \end{pmatrix} + \left[\frac{\Lambda/L}{w_1 - w_2} - \frac{\Lambda/L}{w_2 - w_0}\right] \begin{pmatrix} -1 + H - J \\ \\ J - H \\ \\ 1 \end{pmatrix} \Rightarrow \frac{\partial\lambda_2}{\partial Q} < 0, \frac{\partial w_0}{\partial Q} > 0, \frac{\partial w_1}{\partial Q} > 0, \frac{\partial w_2}{\partial Q} < 0$$

where $H = \frac{\partial g_1}{\lambda_A} = \frac{\partial g_1}{\lambda_C} = -\frac{(Q_d - Q_a)\Lambda\beta}{(w_1 - w_2)^2} < 0$, $J = \frac{\partial g_1}{\lambda_B} = \frac{(Q_d - Q_a)\Lambda\alpha}{(w_1 - w_2)^2} > 0$, $K = \frac{\partial g_2}{\lambda_A} = -\frac{\partial g_2}{\lambda_B} = \frac{\partial g_2}{\lambda_C} = \frac{(Q_a - Q_d + p)\Lambda\alpha}{(w_2 - w_0)^2} > 0$, $L = 1 + 2K + J - H > 0$.

In the case when $Q_a > Q_d$, we have $g_1 = \frac{Q_a - Q_d}{w_2 - w_1}\Lambda$, $g_2 = \frac{p}{w_1 - w_0}\Lambda$, $\lambda_A = \lambda_2$, $\lambda_B = \lambda_1$ and $\lambda_C = \lambda_0$. The comparative statics are:

$$\begin{pmatrix} \partial\lambda_0/\partial p \\ \\ \partial\lambda_1/\partial p \\ \\ \partial\lambda_2/\partial p \end{pmatrix} = \frac{\Lambda/L}{w_1 - w_0}\begin{pmatrix} -1+H-J \\ \\ 1-H+J \\ \\ 0 \end{pmatrix} \Rightarrow \begin{pmatrix} \partial\lambda_0/\partial p < 0 \\ \\ \partial\lambda_1/\partial p > 0 \\ \\ \partial\lambda_2/\partial p = 0 \end{pmatrix} \Rightarrow \frac{\partial w_0}{\partial p} = 0, \frac{\partial w_1}{\partial p} = 0$$

$$\begin{pmatrix} \partial\lambda_0/\partial Q \\ \\ \partial\lambda_1/\partial Q \\ \\ \partial\lambda_2/\partial Q \end{pmatrix} = \frac{\Lambda/L}{w_2 - w_1}\begin{pmatrix} -1+H-J+K \\ \\ -H+J-K \\ \\ 1 \end{pmatrix} \Rightarrow \begin{pmatrix} \partial\lambda_0/\partial Q < 0 \\ \\ \partial\lambda_1/\partial Q > 0 \\ \\ \partial\lambda_2/\partial Q > 0 \end{pmatrix} \Rightarrow \frac{\partial w_0}{\partial Q} < 0, \frac{\partial w_1}{\partial Q} < 0$$

where $H = \frac{\partial g_1}{\lambda_A} = -\frac{(Q_a - Q_d)\Lambda\alpha}{(w_2 - w_1)^2} < 0$, $J = \frac{\partial g_1}{\lambda_B} = \frac{\partial g_1}{\lambda_C} = \frac{(Q_a - Q_d)\Lambda\beta}{(w_2 - w_1)^2} > 0$, $K = \frac{\partial g_2}{\lambda_B} = \frac{\partial g_2}{\lambda_C} = \frac{(Q_a - Q_d)\Lambda(\alpha - \beta)}{(w_1 - w_0)^2} < 0$, $L = 1 + J - H > 0$.

B.5 An Example of Multiple Equilibria When Physician Dual Practice Is Allowed and There Is Outside Service

In the following discussion, we assume that physician dual practice is allowed, the total patient arrival rate is $\Lambda = 1.3501$, $n_d = n_a = 1$, the queue of dual-practice physicians is an $M/G/n_d$ queue with non-preemptive priority, the queue of public-only physicians is an $M/G/n_a$ queue under FIFO service principle, service times of both types of physician follow the same distribution, the first and second moments of service time are $m = 1$ and $m^2 = 1$, the service qualities are $Q_d = 2.2$ and $Q_a = 2$, price $p = 0.45$, and patient's time cost has the following discrete distribution:

$$f(x) = \begin{cases} \frac{25}{135.01}, & x = 0; \\ \frac{32}{135.01}, & x = \frac{1}{4}; \\ \frac{35}{135.01}, & x = \frac{1}{3}; \\ \frac{27}{135.01}, & x = \frac{5}{8}; \\ \frac{16}{135.01}, & x = 1; \\ \frac{0.01}{135.01}, & x = 2. \end{cases}$$

Given the above assumptions and no outside service, one equilibrium is as follows:

Table B.1: Equilibrium without outside service

	h	λ	w
class 0	1, 2	0.1601	0.4346
class 1	0, $\frac{1}{4}$	0.57	1.6104
class 2	$\frac{1}{3}$, $\frac{5}{8}$	0.62	0.8158

Table B.1 shows the composition of each patient class, the arrival rate to each queue and the waiting time of each patient class. For instance, the first row of Table B.1 says that class 0 patients consist of patients with time costs $h = 1$ and $h = 2$, the arrival rate to queue 0 is $\lambda_0 = 0.1601$, and the waiting time of class 0 patients is $w_0 = 0.4346$. Other rows of the table can be interpreted accordingly. Now suppose that patients are offered an outside service which has a reservation net benefit $v = 1$, then patients with time cost $h = 2$ would switch from the existing system to the outside service. The rest of the patients would remain being served in the existing system. In this case, two possible equilibria emerge:

Table B.2: Equilibrium with outside service

	h	λ	w
class 0	1	0.16	0.4345
class 1	$0, \frac{1}{4}$	0.57	1.6093
class 2	$\frac{1}{3}, \frac{5}{8}$	0.62	0.8158
outside service	2	0.01	N/A

Table B.3: Equilibrium with outside service

	h	λ	w
class 0	$\frac{5}{8}, 1$	0.43	0.5965
class 1	0	0.25	1.8640
class 2	$\frac{1}{4}, \frac{1}{3}$	0.67	1.0151
outside service	2	0.01	N/A

Every patient class in the equilibrium of Table B.2 has a waiting time no longer than that in the equilibrium of Table B.1. On the contrary, every patient class in the equilibrium of Table B.3 has a waiting time longer than that in the equilibrium of Table B.1.

We can also compare the patient welfare of the three equilibria as in Table B.4. For instance, a patient with time cost $h = \frac{1}{4}$ has a net benefit of 1.7974 in the equilibrium of Table B.1, a net benefit of 1.7976 in the equilibrium of Table B.2 and a net benefit of 1.7462 in the equilibrium of Table B.3. From Table B.4, we can see that except for patients with time cost $h = 2$, every other type of patients have the highest net benefit in the equilibrium of Table B.2, followed by the net benefit in the equilibrium of Table B.1, and the lowest net benefit in the equilibrium of Table B.3.

Table B.4: Patient welfare comparison

h	Table B.1	Table B.2	Table B.3
0	2.2	2.2	2.2
$\frac{1}{4}$	1.7974	1.7976	1.7462
$\frac{1}{3}$	1.7281	1.7281	1.6616
$\frac{5}{8}$	1.4901	1.4901	1.3772
1	1.3154	1.3155	1.1535
2	0.8808	1	1

Appendix C

Appendix for Chapter 4

C.1 Proofs

C.1.1 Proof of Proposition 4.1

Proof: The proof for the first part is as follows. By (4.2), $\frac{\partial V}{\partial p}$ can be rewritten as:

$$\frac{\partial V}{\partial p} = \frac{1}{\lambda}\int_0^{\lambda} \theta w(\theta)d\theta \left\{1 - \frac{p}{\lambda}\frac{\partial \lambda}{\partial p}\right\} + \lambda + p\frac{\partial \lambda}{\partial p} - \int_0^{\lambda} w(\theta)d\theta - (h-q)\frac{\partial \lambda}{\partial p} \quad \text{(C.1)}$$

We have the following results:

- $\lim_{p\to 0_+} \frac{1}{\lambda}\int_0^{\lambda} \theta w(\theta)d\theta = \lim_{\lambda\to 0_+} \frac{1}{\lambda}\int_0^{\lambda} \theta w(\theta)d\theta = \lim_{\lambda\to 0_+} \lambda \cdot w(\lambda) = 0$ (by *L'Hopital*'s rule);
- $\frac{p}{\lambda}\frac{\partial \lambda}{\partial p} \leq \frac{1}{2}$;
- $\lim_{p\to 0_+} p\frac{\partial \lambda}{\partial p} = 0$;
- $\lim_{p\to 0_+} \int_0^{\lambda} w(\theta)d\theta = \lim_{\lambda\to 0_+} \int_0^{\lambda} w(\theta)d\theta = 0.$

When $p \to 0_+$, the first four items of (C.1) go to 0. Therefore, (C.1) is simplified as:

$$\frac{\partial V}{\partial p}|_{p\to 0_+} = -\lim_{p\to 0_+}(h-q)\frac{\partial \lambda}{\partial p}\begin{cases} <0, & h>q; \\ \\ =0, & h=q. \end{cases}$$

The proof for the second part is as follows. If $\mu < \frac{1+\sqrt{1+4/R}}{2}$, then $\left(\frac{1}{\mu-1}-\frac{1}{\mu}\right) > R$. In this case, the the public health system alone cannot accommodate all patients. Suppose not, then the patient type with $\theta = 1$ would have negative net health benefit, i.e. $R - \left(\frac{1}{\mu-1}-\frac{1}{\mu}\right) < 0$. For any $p \geq R$, no patients would choose private care. Therefore, the effective range of p for the health planner's decision making is $p \in [0,R]$.

If $\mu \geq \frac{1+\sqrt{1+4/R}}{2}$, then $\left(\frac{1}{\mu-1}-\frac{1}{\mu}\right) \leq R$. In this case, the public health system alone can accommodate all patients and provides them with nonnegative net health benefits. For any $p \geq \left(\frac{1}{\mu-1}-\frac{1}{\mu}\right)$, no patients would choose private care and all patients would stay in the public system. Therefore the effective range of p for the health planner's decision making is $p \in \left[0, \left(\frac{1}{\mu-1}-\frac{1}{\mu}\right)\right]$.

In summary, the upper bound of effective p for the health planner's decision making is $p_{\max} = \min\left\{R, \left(\frac{1}{\mu-1}-\frac{1}{\mu}\right)\right\}$. When $p = p_{\max}$, $\lambda = \mathbf{1}$ and the first order derivative of V with respect to p is:

$$\frac{\partial V}{\partial p}|_{p_{\max}} = \int_0^1 \left[1 - p\frac{\partial \lambda}{\partial p}\right]\theta w(\theta) d\theta + (p-h+q)\cdot\frac{\partial \lambda}{\partial p}$$

Since $\left[1-p\frac{\partial \lambda}{\partial p}\right] > 0$ and $(p-h+q) \geq (R-h+q) > 0$, we have $\frac{\partial V}{\partial p}|_{p_{\max}} > 0$.

The proof for the third part is as follows. If $\forall p \in [0, p_{\max}]$, we have $\frac{\partial V}{\partial p} \leq 0$, then $p^* = p_{\max}$. The corresponding tax/subsidy is $t^* = p^* - h = p_{\max} - h$. Otherwise, if there exists a $\widetilde{p} \in [0, p_{\max})$, such that $\frac{\partial V}{\partial p}|_{\widetilde{p}} > 0$. By the continuity of $\frac{\partial V}{\partial p}$ in $(0, \widetilde{p})$, there must exist a $\widehat{p} \in (0, \widetilde{p})$, such that $\frac{\partial V}{\partial p}|_{\widehat{p}} = 0$. Therefore, p^* is either equal to $p_{\max}$ or determined by FOC solutions. ■

C.1.2 Proof of Proposition 4.2

Proof: The idea of the proof is as follows. Since we know that $\frac{\partial V}{\partial p}|_{p\to 0_+} \leq 0$, if $\frac{\partial V}{\partial p}$ is quasi-convex in p, i.e. if we can prove that $\frac{\partial^2 V}{\partial p^2} = 0$ only has one unique solution, then the uniqueness of $\frac{\partial V}{\partial p} = 0$ is established. For ease of exposition, we use α, α'_p and

α''_{pp} to denote $\alpha(p)$, $\frac{\partial \alpha(p)}{\partial p}$ and $\frac{\partial^2 \alpha(p)}{\partial p^2}$ respectively. To derive the sufficient conditions for the quasi-convexity of $\frac{\partial V}{\partial p}$, we start by looking at $\frac{\partial^2 V}{\partial p^2}$:

$$\frac{\partial^2 V}{\partial p^2} = \underbrace{2[\gamma_1(p)+\gamma_2(p)]\mu}_{f_1(p)} - \underbrace{\left[\gamma_1(p)\frac{2}{\lambda^2}\int_0^{\lambda} \theta \cdot w(\theta)d\theta + \gamma_2(p)w(\lambda)\right]\mu}_{f_2(p,\mu)} - (h-q)\underbrace{\alpha''_{pp}\mu}_{f_3(p)} \tag{C.2}$$

where $\gamma_1(p) = \alpha'_p - \gamma_2(p) + \frac{1}{2}p\alpha''_{pp}$, $\gamma_2(p) = p\frac{(\alpha'_p)^2}{\alpha}$. The structure of $\gamma_1(p)$ and $\gamma_2(p)$ is usually messy and it is not easy to see their functional properties. We resort to numerical tests by varying p from 0_+ to a very large positive number to test the convexity/concavity of these functions. These numerical tests show the following results:

(i) For $p \in [0, 10^4]$, both $\gamma_1(p)$ and $\gamma_2(p)$ are positive, strictly decreasing and convex in p. Both function decreases from $+\infty$ to 0_+.

(ii) For $p \in [0, 5000]$, function $\frac{\gamma_1(p)}{\gamma_2(p)}$ is positive, strictly increasing and concave in p. The function increases from 0.5 to 1_-.

(iii) Function $f_3(p)$ is strictly increasing and concave in p. The function increases from from $-\infty$ and to 0_-.

(iv) Function $-\frac{\gamma_1(p)+\gamma_2(p)}{f_3(p)\cdot\alpha^2}$ is strictly decreasing and convex in p. The function decreases from 1.5_- to 1_+.

In the rest of this chapter, we assume $p^{\max} < 10^4$ to preserve the functional properties (i), (ii), (iii) and (iv). When $p = p^{\max}$, $\lambda = 1$, $w(1) < 2$ and $2\int_0^1 \theta w(\theta)d\theta < 2$, so we have $\frac{\partial^2 V}{\partial p^2}|_{p^{\max}} > 0$.

For part 1 of Proposition 4.2, λ is increasing and concave in p, given that $w(\lambda)$ and $\frac{1}{\lambda^2}\int_0^{\lambda} \theta \cdot w(\theta)d\theta$ are convex in λ, $w(\lambda)$ and $\frac{1}{\lambda^2}\int_0^{\lambda} \theta \cdot w(\theta)d\theta$ are decreasing and convex in p. Let $f_1(p)$, $f_2(p)$ and $f_3(p)$ be the functions denoted in (C.2). According to the functional properties (i), (ii), (iii) and (iv), $[f_1(p) - (h-q)f_3(p)]$ is decreasing and convex in p. Because the product of two decreasing and convex functions of p is also decreasing and convex in p, $f_2(p)$ is decreasing and convex in p as well.

Since $\frac{\partial^2 V}{\partial p^2}|_{p\to 0_+} < 0$ and $\frac{\partial^2 V}{\partial p^2}|_{p^{\max}} > 0$, we have $[f_1(0_+)-(h-q)f_3(0_+)] - f_2(0_+) < 0$ and $[f_1(p^{\max}) - (h-q)f_3(p^{\max})] - f_2(p^{\max}) > 0$. By the convexity and monotonicity of $[f_1(p)-(h-q)f_3(p)]$ and $f_2(p)$, $f_1(p)-(h-q)f_3(p) = f_2(p)$ has one and only one solution. Let this solution be denoted as $\hat{p}$. Therefore, when $p \in [0, \hat{p}]$, we have $\frac{\partial^2 V}{\partial p^2} < 0$; when $p \in [\hat{p}, p_{\max}]$, we have $\frac{\partial^2 V}{\partial p^2} > 0$.

Since $\lim_{p \to 0_+} \frac{\partial V}{\partial p} \leq 0$ and $\lim_{p \to p^{\max}} \frac{\partial V}{\partial p} > 0$, when p goes from 0 to $\hat{p}$, $\frac{\partial V}{\partial p}$ goes from 0 to the lowest point (negative). When p goes from $\hat{p}$ to $p_{\max}$, $\frac{\partial V}{\partial p}$ goes from the lowest point (negative) to the highest point (positive). Therefore, $\frac{\partial V}{\partial p}$ is a quasi-convex function of p. Using the similar logic, $V(p|\mu)$ is a quasi-convex function of p and there must exist a unique point $p^* \in [\hat{p}, p^{\max}]$ such that $\frac{\partial V}{\partial p}|_{p^*} = 0$, with p^* being the optimal price.

For part 2 of Proposition 4.2, we have

$$\frac{2}{\lambda^2} \int_0^{\lambda} \theta \cdot w(\theta) d\theta - w(\lambda) = \frac{2}{\lambda^2} \int_0^{\lambda} \theta [w(\theta) - w(\lambda)] d\theta > 0$$

Given that $2 + \frac{2}{3}(h-q)\alpha^{-2} \geq \frac{2}{\lambda^2}{}^{\lambda}_{0} \theta \cdot w(\theta) d\theta$, $\forall p \in R^+$, by (C.2), we know that

$$\begin{aligned} & \frac{\partial^2 V}{\partial p^2} \\ \geq \;& 2[\gamma_1(p) + \gamma_2(p)] - \frac{2}{\lambda^2} \int_0^{\lambda} \theta \cdot w(\theta) d\theta [\gamma_1(p) + \gamma_2(p)] + \frac{2}{3}(h-q)\alpha^{-2}[\gamma_1(p) + \gamma_2(p)] \\ = \;& \left\{ 2 - \frac{2}{\lambda^2} \int_0^{\lambda} \theta \cdot w(\theta) d\theta + \frac{2}{3}(h-q)\alpha^{-2} \right\} [\gamma_1(p) + \gamma_2(p)] \geq 0 \end{aligned}$$

In this case, $\frac{\partial V}{\partial p}$ is increasing in p. Since $\lim_{p \to 0_+} \frac{\partial V}{\partial p} \leq 0$ and $\lim_{p \to p^{\max}} \frac{\partial V}{\partial p} > 0$, by the continuity of the function, there must exist one unique point p^* such that $\frac{\partial V}{\partial p}|_{p^*} = 0$, with p^* being the optimal price. ■

C.1.3 Proof of Proposition 4.3

Proof: Let $\lambda_1^* = \alpha(p_1^*)\mu$ and $\lambda_2^* = \alpha(p_2^*)\mu$, then $p_1^* > p_2^*$ if and only if $\lambda_1^* > \lambda_2^*$. We evaluate the first-order derivatives at p_2^* and we can assert that:

$$\frac{\partial V}{\partial p}|_{\{w_1(\theta), p_2^*\}} - \frac{\partial V}{\partial p}|_{\{w_2(\theta), p_2^*\}} = \frac{\partial V}{\partial p}|_{\{w_1(\theta), p_2^*\}} = \int_0^{\lambda_2^*} g(\theta)[w_1(\theta) - w_2(\theta)] d\theta < 0 \tag{C.3}$$

where $g(\theta) = \left\{ \frac{\theta}{\lambda_2^*} \left[1 - \frac{p_2^*}{\lambda_2^*} \cdot \frac{\partial \lambda}{\partial p}|_{p_2^*} \right] - 1 \right\}$ and $g(\theta)$ is increasing in θ. The proof is as follows.

First, we have $\frac{\theta}{\lambda_2^*} < 1$ and $\frac{p_2^*}{\lambda_2^*}\frac{\partial\lambda}{\partial p}|_{p_2^*} \leq 0.5$, so it must be that $g(\theta) < 0, \forall\theta \in [0, \lambda_2^*]$. If $w_1(\theta) - w_2(\theta) > 0, \forall\theta \in [0, \lambda_2^*]$, then (C.3) holds. If not, $[0, \lambda_2^*]$ must be divided into alternative segments: $[0, y_1]$, $[y_1, y_2]$, ..., $[y_n, \lambda_2^*]$ such that

$$w_1(\theta) - w_2(\theta) \begin{cases} > 0 & \text{if } \theta \in [y_{2k}, y_{2k+1}] \\ < 0 & \text{if } \theta \in [y_{2k+1}, y_{2k+2}] \end{cases}$$

where $k \in \mathbb{N}$. $\forall\theta \in [y_{2k}, y_{2k+1}], g(\theta)[w_1(\theta) - w_2(\theta)] < g(y_{2k+1})[w_1(\theta) - w_2(\theta)] < 0$. $\forall\theta \in [y_{2k+1}, y_{2k+2}], 0 < g(\theta)[w_1(\theta) - w_2(\theta)] < g(y_{2k+1})[w_1(\theta) - w_2(\theta)]$. Therefore, we always have

$$\int_{y_{2k}}^{y_{2k+2}} g(\theta)[w_1(\theta) - w_2(\theta)]d\theta < g(y_{2k+1})\int_{y_{2k}}^{y_{2k+2}} [w_1(\theta) - w_2(\theta)]d\theta$$

For $k = 0$, we have

$$\int_0^{y_2} g(\theta)[w_1(\theta) - w_2(\theta)]d\theta < g(y_1)\int_0^{y_2} [w_1(\theta) - w_2(\theta)]d\theta < 0$$

Since $g(y_1) < g(y_3) < 0$, we have

$$g(y_1)\int_{y_0}^{y_2}[w_1(\theta)-w_2(\theta)]d\theta + g(y_3)\int_{y_2}^{y_4}[w_1(\theta)-w_2(\theta)]d\theta < g(y_3)\int_{y_0}^{y_4}[w_1(\theta)-w_2(\theta)]d\theta < 0$$

Therefore, we have

$$\int_0^{y_4} g(\theta)[w_1(\theta) - w_2(\theta)]d\theta < g(y_3)\int_0^{y_4} [w_1(\theta) - w_2(\theta)]d\theta$$

By the same logic, we can prove that for any $k \in \mathbb{N}$

$$\int_0^{y_{2k+2}} g(\theta)[w_1(\theta) - w_2(\theta)]d\theta < g(y_{2k+1})\int_0^{y_{2k+2}} [w_1(\theta) - w_2(\theta)]d\theta$$

If $n = 2i + 1$, then we have

$$\int_0^{\lambda_2^*} g(\theta)[w_1(\theta) - w_2(\theta)]d\theta < g(y_{2k+1})\int_0^{y_{2k+2}} [w_1(\theta) - w_2(\theta)]d\theta < 0$$

If $n = 2i+2$, then we have

$$\begin{aligned} & \int_0^{\lambda_2^*} g(\theta)[w_1(\theta)-w_2(\theta)]d\theta \\ < \quad & g(y_{2k+1})\int_0^{y_{2k+2}}[w_1(\theta)-w_2(\theta)]d\theta + \int_{y_{2k+2}}^{\lambda_2^*} g(\theta)[w_1(\theta)-w_2(\theta)]d\theta < 0 \end{aligned}$$

which follows that $w_1(\theta)-w_2(\theta) > 0$, $g(\theta) < 0$, $\forall\theta \in [y_{2k+2}, \lambda_2^*]$. Because $\frac{\partial V}{\partial p}|_{w_1(\theta)} = 0$ only has one solution and $\frac{\partial V}{\partial p}|_{w_1(\theta),p_2^*} < 0$, p_1^* must be on the right-hand side of p_2^*, i.e. $p_2^* > p_1^*$. ■

C.1.4 Proof of Proposition 4.4

Proof: Let $\tilde{p}_1$ and $\tilde{p}_2$ be such that $\frac{\partial V}{\partial p}|_{\{\mu_1,\tilde{p}_1\}} = \frac{\partial V}{\partial p}|_{\{\mu_2,\tilde{p}_2\}} = 0$, where $\mu_1 < \mu_2$. To prove the theorem, we only need to prove $\tilde{p}_1 > \tilde{p}_2$. Since both $\frac{1}{\lambda^2}\int_0^{\lambda}\theta\cdot w(\theta)d\theta$ and $w(\lambda)$ are decreasing in λ, and $\lambda = \alpha\cdot\mu$, we know that $f_2(p,\mu)$ in (C.2) is decreasing in μ. We must have $\int_0^{\tilde{p}}[f_1(p) - f_2(p,\mu) - (h-q)f_3(p)]dp$ increasing in μ, $\forall\tilde{p} > 0$. According to (C.2), we have

$$\begin{aligned} & \frac{\partial V}{\partial p}|_{\{\mu_1,\tilde{p}_1\}} = \lim_{p\to 0_+}\frac{\partial V}{\partial p}|_{\mu_1} + \int_0^{\tilde{p}_1}\frac{\partial^2 V}{\partial p^2}|_{\mu_1}dp \\ & = \mu_1\left\{-(h-q)\lim_{p\to 0_+}\alpha' + \int_0^{\tilde{p}_1}[f_1(p) - f_2(p,\mu_1) - (h-q)f_3(p)]\,dp\right\} = 0 \end{aligned} \tag{C.4}$$

It must be that the bracket in (C.4) equal to zero. Similarly, we have

$$\begin{aligned} & \frac{\partial V}{\partial p}|_{\{\mu_2,\tilde{p}_1\}} = \lim_{p\to 0_+}\frac{\partial V}{\partial p}|_{\mu_2} + \int_0^{\tilde{p}_1}\frac{\partial^2 V}{\partial p^2}|_{\mu_2}dp \\ & = \mu_2\left\{-(h-q)\lim_{p\to 0_+}\alpha' + \int_0^{\tilde{p}_1}[f_1(p) - f_2(p,\mu_2) - (h-q)f_3(p)]\,dp\right\} \end{aligned} \tag{C.5}$$

Since $\mu_1 < \mu_2$ and $\int_0^{\tilde{p}_1}\{f_1(p) - f_2(p,\mu) - (h-q)f_3(p)\}dp$ is increasing in μ, we know that the bracket in (C.5) is greater than zero, so $\frac{\partial V}{\partial p}|_{\{\mu_2,\tilde{p}_1\}} > 0$. Since $\frac{\partial V}{\partial p}|_{\mu_2} = 0$ only has one unique solution, $\tilde{p}_2$ must be on the left hand side of $\tilde{p}_1$, so we have $\tilde{p}_2 < \tilde{p}_1$. ■

C.1.5 Proof of Proposition 4.5

Proof: To find out μ_w, we only need to input $p^* = h$ to the first order condition and solve for μ. Given a welfare weight function $w(\theta)$, let $\hat{\lambda} = \alpha(h) \cdot \mu$. Substituting $\hat{\lambda}$ for λ in (4.2) gives rise to:

$$\int_0^{\hat{\lambda}} \left(1 - \frac{\theta}{\hat{\lambda}} + \frac{\theta \cdot h}{(\hat{\lambda})^2} \frac{\partial \hat{\lambda}}{\partial h}\right) w(\theta) d\theta = [\alpha(h) + q\alpha'(h)]\mu \tag{C.6}$$

Solving equation (C.6) for μ yields μ_w, which is a function of $w(\theta)$ and h. If $w_1(\theta)$ stochastically dominates $w_2(\theta)$, by the proof of Proposition 4.3, we have

$$[\alpha(h) + q\alpha'(h)]\,(\mu_{w_1} - \mu_{w_2}) = \int_0^{\hat{\lambda}} \left(1 - \frac{\theta}{\hat{\lambda}} + \theta \frac{h}{(\hat{\lambda})^2} \frac{\partial \hat{\lambda}}{\partial h}\right) [w_1(\theta) - w_2(\theta)] d\theta \geq 0$$

Therefore, we must have $\mu_{w_1} \geq \mu_{w_2}$. ■

C.1.6 Proof of Proposition 4.6

Proof: Taking the first order derivative of $J(\mu)$ with respect to μ gives rise to:

$$\frac{\partial J(\mu)}{\partial \mu} = \alpha_h \left\{ q - \frac{h}{\widetilde{\lambda}^2} \int_0^{\widetilde{\lambda}} \theta \cdot w(\theta) d\theta \right\} - p^* \alpha^* \left\{ 1 - \frac{1}{\lambda^{*2}} \int_0^{\lambda^*} \theta \cdot w(\theta) d\theta \right\} + (h - q)\alpha^* \tag{C.7}$$

where $\alpha_h = \alpha(h)$, $\alpha^* = \alpha(p^*(\mu))$, $\widetilde{\lambda} = \alpha_h \cdot \mu$ and $\lambda^* = \alpha^* \mu$. Substituting $(1 - \sigma)\theta^{-\sigma}$ for $w(\theta)$ in FOC gives rise to:

$$\lambda^\sigma = (\alpha\mu)^\sigma = \frac{1 + (1 - \sigma) p \frac{\alpha'}{\alpha}}{(2 - \sigma) \left(1 + (p - h + q) \frac{\alpha'}{\alpha}\right)}$$

$$\Rightarrow \quad \mu^{-\sigma} = \alpha^\sigma (2 - \sigma) \frac{1 + p \frac{\alpha'}{\alpha}}{1 + (1 - \sigma) p \frac{\alpha'}{\alpha}} \tag{C.8}$$

Inputting $w(\theta) = (1 - \sigma)\theta^{-\sigma}$ and (C.8) into (C.7) yields:

$$
\begin{aligned}
\frac{\partial J(\mu)}{\partial \mu} &= \alpha_h\left\{q-h\frac{1-\sigma}{2-\sigma}\tilde{\lambda}^{-\sigma}\right\}-p^*\alpha^*\left\{1-\frac{1-\sigma}{2-\sigma}\lambda^{*-\sigma}\right\}+(h-q)\alpha^* \\
&= (1-\sigma)(\alpha^*)^{\sigma}\left[p^*\cdot(\alpha^*)^{1-\sigma}-h\cdot(\alpha_h)^{1-\sigma}\right]\frac{1+(p^*-h+q)\frac{\alpha^{*\prime}}{\alpha^*}}{1+(1-\sigma)p^*\frac{\alpha^{*\prime}}{\alpha^*}} \\
&\quad -\left[(p^*-h+q)\alpha^*-q\cdot\alpha_h\right] \qquad \text{(C.9)}
\end{aligned}
$$

By (C.9), $\frac{\partial J(\mu)}{\partial \mu}$ is a function of p^* only. Numerical tests show that:

$$
\frac{\partial J(\mu)}{\partial \mu}\begin{cases} >0, & \text{if } p^*<h; \\ =0, & \text{if } p^*=h; \\ <0, & \text{if } p^*>h. \end{cases}
$$

Since $p^*(\mu)$ is decreasing in μ, when $\mu \leq \mu_w$, we have $p^* > h$. In this case $\frac{\partial J(\mu)}{\partial \mu} < 0$, i.e. the marginal improvement using the optimal tax/subsidy is negative, so less improvement is obtained when μ increases. When $p^* = h$, there is no improvement. When $\mu > \mu_w$, we have $p^* < h$. In this case $\frac{\partial J(\mu)}{\partial \mu} > 0$, i.e. the marginal improvement using the optimal tax/subsidy is positive, so more improvement will be obtained when μ increases. ■

C.1.7 Proof of Proposition 4.7

Proof: Let $\phi(\lambda) = 1 - \frac{1}{\lambda^2}\int_0^{\lambda} w(\theta)\theta d\theta$, then

$$
\phi' = \frac{2}{\lambda^3}\int_0^{\lambda}\theta\cdot w(\theta)d\theta - \frac{w(\lambda)}{\lambda} = \frac{2}{\lambda^3}\int_0^{\lambda}\theta\cdot[w(\theta)-w(\lambda)]d\theta > 0
$$

By (4.4), $\frac{\partial V}{\partial \lambda} = \frac{c}{\alpha} + q - h + p\cdot\phi(\lambda)$ is an increasing function of λ with $\lambda_{\min} = 0_+$ and $\lambda_{\max} = 1_-$.

$$
\begin{aligned}
\frac{\partial V}{\partial \lambda}|_{\lambda=0_+} &= \frac{c}{\alpha}+q+(p-h)-p\frac{w(0_+)}{2} \\
\frac{\partial V}{\partial \lambda}|_{\lambda=1_-} &= \frac{c}{\alpha}+q+(p-h)-p\int_0^1\theta\cdot w(\theta)d\theta
\end{aligned}
$$

The two equations above follow because when $\lambda \to 0$, we have $\lambda^2 \to 0$, $\int_0^{\lambda} w(\theta)\theta d\theta \to$

0 (according to assumption $\lim_{\theta \to 0_+} w(\theta) \cdot \theta = 0$). By *L Ĥopital's rule*, we have

$$\lim_{\lambda \to 0} \frac{1}{\lambda^2} \int_0^{\lambda} w(\theta)\theta d\theta = \lim_{\lambda \to 0} \frac{\lambda w(\lambda)}{2\lambda} = \frac{w(0_+)}{2}$$

If $\frac{\partial V}{\partial \lambda}|_{\lambda=0_+} \geq 0$, then $\frac{\partial V}{\partial \lambda} \geq 0, \forall \lambda \in [0,1]$, so we have $\lambda^* = 0$ and $\mu^* = \frac{\lambda^*}{\alpha} = 0$. If $\frac{\partial V}{\partial \lambda}|_{\lambda=1_-} \leq 0$, then $\frac{\partial V}{\partial \lambda} \leq 0, \forall \lambda \in [0,1]$, so we have $\lambda^* = 1$ and $\mu^* = \frac{\lambda^*}{\alpha} = \frac{1}{\alpha}$. Otherwise, λ^* is determined by $\frac{\partial V}{\partial \lambda} = 0$. ■

C.1.8 Proof of Proposition 4.8

Proof: If $w(\theta) = 1$, (4.2) becomes

$$\frac{\partial V}{\partial p} = \frac{1}{2}\left(p \cdot \frac{\partial \lambda}{\partial p} + \lambda\right) - (h-q) \cdot \frac{\partial \lambda}{\partial p}$$

We know that $\left(p \cdot \frac{\partial \lambda}{\partial p} + \lambda\right)$ is strictly increasing in p. The value of $\left(p \cdot \frac{\partial \lambda}{\partial p} + \lambda\right)$ goes from 0_+ to 1 when p goes from 0 to $+\infty$. $\frac{\partial \lambda}{\partial p}$ is strictly decreasing in p. The value of $\frac{\partial \lambda}{\partial p}$ goes from $+\infty$ to 0_+ when p goes from 0 to $+\infty$. Therefore $\frac{\partial V}{\partial p} = 0$ has a unique solution and the optimal p^* (thus t^*) is determined by FOC solution. We further have

$$\frac{\partial V}{\partial p} = 0 \quad \Rightarrow \quad \underbrace{\frac{1}{2}\left(p + \frac{\alpha}{\alpha'_p}\right)|_{p=p^*}}_{f(p^*)} = (h-q) \tag{C.10}$$

The left hand side of (C.10) is a function of p^*, denoted as $f(p^*)$. We have $f'(p^*) > 0$, $f''(p^*) > 0$ and $p^* = f^{-1}(h-q)$. If $h = q$, then we have $f(p^*) = 0 \Rightarrow p^* = 0$ and thus $t^* = p^* - q = -q$.

If $h > q$, then $0 < \frac{\partial p^*}{\partial h} = \frac{1}{f'} < \frac{2}{3}$ and $\frac{\partial^2 p^*}{\partial h^2} = -\frac{f''}{(f')^3} < 0$, so $p^* = q + \int_q^h \frac{\partial p^*}{\partial h} dh > 0$. As $p^* = h + t^* > 0$, we have $t^* > -h$ and $\frac{\partial t^*}{\partial h} = \frac{1}{f'} - 1 < 0$. Therefore, when $h > q$, we must have $t^* = \int_q^h \frac{\partial t^*}{\partial h} dh - q < 0$. ■

C.1.9 Proof of Proposition 4.10

Proof: According to FOC solution (C.10), we have

$$\frac{p^*}{2}-(h-q)=-\frac{\alpha}{2\alpha'}|_{p=p^*}$$

Because $\frac{\alpha}{2\alpha'}$ is strictly increasing in p, and p^* is strictly increasing in h, so $\frac{\alpha}{2\alpha'}|_{p^*}$ is strictly increasing in h. (4.7) can be rewritten as

$$\frac{\partial V}{\partial \mu}|_{p^*(\mu)}=c-\frac{\alpha^2}{2\alpha'}|_{p^*} \tag{C.11}$$

Because (C.11) is strictly decreasing in h, there must exist h^* that solves $c-\frac{\alpha^2}{2\alpha'}|_{p^*}=0$. If $h\leq h^*$, $\frac{\partial V}{\partial \mu}|_{p^*(\mu)}\geq 0,\forall\mu\geq 0$, so it is optimal to set $\mu^*=0$. In this case, all patients choose private care and the total cost is $V=h\cdot\mathbf{1}$. If $h>h^*$, $\frac{\partial V}{\partial \mu}|_{p^*(\mu)}<0,\forall\mu\geq 0$, so it is optimal to set μ^* that solves $\frac{1}{\mu-1}-\frac{1}{\mu}=p^*(h)$. In this case, all patients would be served in the public health system and the total cost is $V=c\mu^*+q+\frac{p^*(h)}{2}$. ■

C.2 Analysis of Proposition 4.4

We use weight function $w(\theta)=(1-\sigma)\theta^{-\sigma},\sigma\in(0,1)$, and the marginal cost V of p can be written as:

$$\frac{\partial V}{\partial p}=\underbrace{\frac{\partial \zeta_1}{\partial p}-\frac{\partial \zeta_2}{\partial p}}_{\text{Marginal Supply-Side Cost}}+\underbrace{\frac{\partial \zeta_3}{\partial p}+\frac{\partial \zeta_4}{\partial p}}_{\text{Marginal Patient-Side Cost}}$$

where

$\frac{\partial \zeta_1}{\partial p}=\alpha' q\mu>0$ $\qquad$ $\frac{\partial \zeta_2}{\partial p}=1-[\alpha+(p-h)\alpha']\mu$

$\frac{\partial \zeta_3}{\partial p}=\frac{1-\sigma}{2-\sigma}[\alpha+(1-\sigma)p\alpha']\alpha^{-\sigma}\mu^{1-\sigma}>0$ $\qquad$ $\frac{\partial \zeta_4}{\partial p}=1-[\alpha+(1-\sigma)p\alpha']\alpha^{-\sigma}\mu^{1-\sigma}$

Both the marginal production cost $\frac{\partial \zeta_1}{\partial p}$ and the marginal waiting cost $\frac{\partial \zeta_3}{\partial p}$ are positive. In order to lower the total cost, the only chance is to make the marginal tax/-subsidy $\frac{\partial \zeta_2}{\partial p}>0$ and/or to make the marginal private care cost $\frac{\partial \zeta_4}{\partial p}<0$, which can be achieved when p is relatively low. When p is relatively low (e.g., close to 0_+), the health planner uses subsidy ($t=p-h<0$), so that $\frac{\partial \zeta_2}{\partial p}\sim O(1+h\alpha'\mu)>0$. In this case, an increase of p means that less amount of subsidy is provided to each private patient, so the health planner ends up paying less total amount of subsidy. When p

is relatively high, e.g., p is close to R_- or $\left(\frac{1}{\mu-1}-\frac{1}{\mu}\right)_-$, the health planner uses tax ($t = p-h > 0$) and we have $(1-\lambda) \to 0_+$, so $\frac{\partial \zeta_2}{\partial p} \sim O(-(p-h)\alpha'\mu) < 0$. In this case, an increase of p means that more tax is charged on each private patient. But since less patients are joining the private system, less total amount of tax revenue is collected from private system.

When we group these marginal costs into “marginal supply-side cost” and “marginal patient-side cost”, the marginal “patient-side” cost is always positive. This is because $[\alpha+(1-\sigma)p\alpha']\,\alpha^{-\sigma}\mu^{1-\sigma} = \lambda^{1-\sigma}[1+(1-\sigma)p\frac{\alpha'}{\alpha}] < \frac{3}{2}, \forall\lambda \in [0,1]$ and $\max\limits_{p\geq 0}\dfrac{p\cdot\alpha'}{\alpha} = 0.5$. We have numerically tested that $\frac{\alpha^{-\sigma}}{2-\sigma}\left[\alpha+(1-\sigma)p\cdot\alpha'\right]\mu^{1-\sigma} < 1, \forall p \in [0, \frac{1}{\mu-1} - \frac{1}{\mu}]$ for a wide range of μ and σ. In this case, the only opportunity to reduce the total cost is to make the marginal “supply-side” cost negative in a greater magnitude than that of the marginal “patient-side” cost. Since $[\alpha+(p-h+q)\alpha']\,\mu$ is increasing in p, we have $\lim\limits_{p\to 0_+}\left[\alpha+(p-h+q)\alpha'\right]\mu - 1 < 0$, and $\lim\limits_{p\to p^{\max}}\left[\alpha+(p-h+q)\alpha'\right]\mu - 1 > 0$, so we can conclude that the marginal “supply-side” cost is negative when p is relatively small.

Now suppose we have μ_1 and $p^*(\mu_1)$ such that $\frac{\partial V}{\partial p}|_{\mu_1,p^*(\mu_1)} = 0$, then we must have $\left(\frac{\partial\zeta_1}{\partial p}-\frac{\partial\zeta_2}{\partial p}\right)|_{\mu_1,p^*(\mu_1)} < 0$, $\left(\frac{\partial\zeta_3}{\partial p}+\frac{\partial\zeta_4}{\partial p}\right)|_{\mu_1,p^*(\mu_1)} > 0$, $\left\|\frac{\partial\zeta_1}{\partial p}-\frac{\partial\zeta_2}{\partial p}\right\|_{\mu_1,p^*(\mu_1)} = \left\|\frac{\partial\zeta_3}{\partial p}+\frac{\partial\zeta_4}{\partial p}\right\|_{\mu_1,p^*(\mu_1)}$. To understand why $p^*(\mu)$ decreases in μ, we look at the second order cross partial derivatives with respect to p and μ. We have the following second order cross partial derivatives:

$$\begin{aligned}
\frac{\partial^2\zeta_1}{\partial p\partial\mu} &= \alpha' q > 0\\
\frac{\partial^2\zeta_2}{\partial p\partial\mu} &= -\left[\alpha+(p-h)\alpha'\right]\\
\frac{\partial^2\zeta_3}{\partial p\partial\mu} &= \frac{(1-\sigma)^2}{2-\sigma}\left[\alpha+(1-\sigma)p\alpha'\right][\alpha\mu]^{-\sigma} > 0\\
\frac{\partial^2\zeta_4}{\partial p\partial\mu} &= -(1-\sigma)\left[\alpha+(1-\sigma)p\alpha'\right][\alpha\mu]^{-\sigma} < 0
\end{aligned}$$

and we know that $\left(\frac{\partial^2\zeta_1}{\partial p\partial\mu}-\frac{\partial^2\zeta_2}{\partial p\partial\mu}\right)|_{\mu_1,p^*(\mu_1)} > 0$, $\left(\frac{\partial^2\zeta_3}{\partial p\partial\mu}+\frac{\partial^2\zeta_4}{\partial p\partial\mu}\right)|_{\mu_1,p^*(\mu_1)} < 0$ and $\left\|\frac{\partial^2\zeta_1}{\partial p\partial\mu}-\frac{\partial^2\zeta_2}{\partial p\partial\mu}\right\|_{\mu_1,p^*(\mu_1)} > \left\|\frac{\partial^2\zeta_3}{\partial p\partial\mu}+\frac{\partial^2\zeta_4}{\partial p\partial\mu}\right\|_{\mu_1,p^*(\mu_1)}$. Therefore, we have $\frac{\partial^2 V}{\partial p\partial\mu}|_{\mu_1,p^*(\mu_1)} >$

0. For any μ_2 such that $\mu_2 > \mu_1$, we must have $\frac{\partial V}{\partial p}|_{\mu_2, p^*(\mu_1)} > 0$. Because $\frac{\partial V}{\partial p} = 0$ only has one unique solution, the FOC solution p_2^* when μ_2 is the public service rate must be on the left hand side of p_1^*.

C.3 Numerical Examples

Table C.1: Tax/subsidy for $M/G/1$ queue with $q = 0.2, c = 0.2, h = 0.5, v = 0.5$.

$\mu \setminus \sigma$	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90
1.10	-0.317	-0.294	-0.262	-0.219	-0.159	-0.075	0.051	0.251	0.613
1.15	-0.318	-0.295	-0.265	-0.226	-0.172	-0.097	0.010	0.173	0.444
1.20	-0.318	-0.296	-0.269	-0.232	-0.183	-0.116	-0.024	0.112	0.326
1.25	-0.318	-0.298	-0.271	-0.237	-0.193	-0.133	-0.051	0.064	0.238
1.30	-0.318	-0.299	-0.274	-0.242	-0.201	-0.147	-0.075	0.025	0.170
1.35	-0.319	-0.300	-0.276	-0.246	-0.208	-0.159	-0.095	-0.007	0.116
1.40	-0.319	-0.301	-0.278	-0.250	-0.215	-0.170	-0.112	-0.035	0.072
1.45	-0.319	-0.301	-0.280	-0.254	-0.221	-0.180	-0.127	-0.058	0.035
1.50	-0.319	-0.302	-0.282	-0.257	-0.226	-0.188	-0.140	-0.078	0.005
1.55	-0.319	-0.303	-0.284	-0.260	-0.231	-0.196	-0.151	-0.095	-0.021
1.60	-0.319	-0.304	-0.285	-0.263	-0.236	-0.203	-0.162	-0.110	-0.044
1.65	-0.319	-0.304	-0.286	-0.265	-0.240	-0.209	-0.171	-0.123	-0.063
1.70	-0.319	-0.305	-0.288	-0.267	-0.243	-0.214	-0.179	-0.135	-0.080
1.75	-0.319	-0.305	-0.289	-0.270	-0.247	-0.219	-0.186	-0.146	-0.095
1.80	-0.319	-0.306	-0.290	-0.271	-0.250	-0.224	-0.193	-0.155	-0.108
1.85	-0.319	-0.306	-0.291	-0.273	-0.253	-0.228	-0.199	-0.164	-0.120
1.90	-0.319	-0.306	-0.292	-0.275	-0.255	-0.232	-0.205	-0.172	-0.131
1.95	-0.319	-0.307	-0.293	-0.276	-0.258	-0.236	-0.210	-0.179	-0.141
2.00	-0.319	-0.307	-0.293	-0.278	-0.260	-0.239	-0.214	-0.185	-0.149
2.05	-0.319	-0.307	-0.294	-0.279	-0.262	-0.242	-0.219	-0.191	-0.157
2.10	-0.319	-0.307	-0.295	-0.280	-0.264	-0.245	-0.222	-0.196	-0.165
2.15	-0.320	-0.308	-0.295	-0.282	-0.266	-0.247	-0.226	-0.201	-0.171
2.20	-0.320	-0.308	-0.296	-0.283	-0.267	-0.250	-0.230	-0.206	-0.177
2.25	-0.320	-0.308	-0.297	-0.284	-0.269	-0.252	-0.233	-0.210	-0.183
2.30	-0.320	-0.308	-0.297	-0.285	-0.270	-0.254	-0.236	-0.214	-0.188
2.35	-0.320	-0.308	-0.298	-0.285	-0.272	-0.256	-0.238	-0.218	-0.193
2.40	-0.320	-0.308	-0.298	-0.286	-0.273	-0.258	-0.241	-0.221	-0.197
2.45	-0.320	-0.309	-0.298	-0.287	-0.274	-0.260	-0.243	-0.224	-0.202
2.50	-0.320	-0.309	-0.299	-0.288	-0.275	-0.261	-0.245	-0.227	-0.205
2.55	-0.320	-0.309	-0.299	-0.288	-0.276	-0.263	-0.248	-0.230	-0.209
2.60	-0.321	-0.309	-0.299	-0.289	-0.277	-0.264	-0.250	-0.232	-0.212
2.65	-0.321	-0.309	-0.300	-0.290	-0.278	-0.266	-0.251	-0.235	-0.216
2.70	-0.321	-0.309	-0.300	-0.290	-0.279	-0.267	-0.253	-0.237	-0.219
2.75	-0.321	-0.309	-0.300	-0.291	-0.280	-0.268	-0.255	-0.239	-0.221
2.80	-0.321	-0.309	-0.301	-0.291	-0.281	-0.269	-0.256	-0.241	-0.224
2.85	-0.321	-0.309	-0.301	-0.292	-0.282	-0.270	-0.258	-0.243	-0.226
2.90	-0.321	-0.309	-0.301	-0.292	-0.282	-0.272	-0.259	-0.245	-0.229
2.95	-0.321	-0.309	-0.301	-0.293	-0.283	-0.273	-0.261	-0.247	-0.231
3.00	-0.322	-0.309	-0.302	-0.293	-0.284	-0.273	-0.262	-0.248	-0.233