ABSTRACT

YAESOUBI, REZA. On the Implementation of Medical Programs in Health Care Systems: Game-Theoretic Frameworks. (Under the direction of Dr. Stephen D. Roberts.)

Finding the cost-effective medical programs, guidelines, and policies have been the major focus of studies in health care resource allocation. There are, however, many medical programs and guidelines that have proven to be cost-effective and to improve the social welfare, but have not been properly *implemented* in the society. Successful implementation of a medical program relies on different factors such as, the health purchasers' willingness to reimburse for the program, the health providers' willingness to offer the program, and the population's willingness to consume the underlying medical program.

This dissertation consists of three papers, each attempts to discuss one or several aspects of medical implementation in a health care system. In the first paper, we develop a game-theoretic framework for estimating a health purchaser's willingness-to-pay (WTP) for health, which is defined as the amount of money the health purchaser (e.g., a health maximizing public agency or a profit maximizing health insurer) is willing to spend for an additional unit of health. We discuss how the *WTP for health* can be employed to determine the medical guidelines, and to price the new medical technologies, such that the health purchaser finds them worthwhile to implement. The framework further introduces a measure for *WTP for expansion*, defined as the amount of money the health purchaser is willing to pay for one percent of increase in the consumption level of an intervention. This measure can be employed to find how much to invest in expanding a medical technology through opening new facilities, advertising, educating the population, etc. Applying our framework to Colorectal Cancer screening tests, we estimate the WTP for health to be $\$9,950$ per quality-adjusted life years, and the WTP for expanding Colonoscopy to be \$45.40 per person per percent increase, for the 2005 U.S. population.

The second paper discusses "coordinating contracts" in a preventive health care system consisting of two noncooperative parties: a health purchaser (e.g., a health insurer) and a health provider (e.g., a hospital). A principal-agent model is proposed to capture the interaction between the two parties. In this model, the health provider determines the type of patients who need to undergo a preventive medical intervention, and get reimbursed by the health purchaser based on the number of patients for whom the intervention is administered. We determine the contracts that *coordinate* the health purchaser-health provider relationship; i.e., the contracts that allow each entity to optimize its own objective function while maximizing the population's welfare. We characterize the coordinating contracts for two settings: we show that under certain conditions (1) when the number of customers for the medical intervention is verifiable, there exist a gate-keeping contract and a set of concave contracts that coordinate the system; and

(2) when the number of customers is not verifiable, contracts of bounded linear and bounded nonlinear forms can coordinate the system.

The notion of coordinating contracts is extended in the third paper to health systems with limited capacity in providing the underlying medical intervention. In the new setting, the health provider should allocate (or build) the medical capacity before observing the demand for the medical intervention. We show that (1) when the number of customers for the medical intervention is verifiable, a piece-wise linear contact can coordinate the system; and (2) when the number of customers is not verifiable, a menu of incentive-feasible piece-wise linear contacts can coordinate the system. We characterize the coordinating contracts under each setting.

On the Implementation of Medical Programs in Health Care Systems: Game-Theoretic Frameworks

by
Reza Yaesoubi

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Industrial Engineering

Raleigh, North Carolina

2010

APPROVED BY:

Dr. Julie S. Ivy	Dr. Michael Pignone
Dr. Stephen D. Roberts Chair of Advisory Committee	Dr. Theofanis Tsoulouhas
Dr. Reha Uzsoy	

BIOGRAPHY

Reza Yaesoubi is a Ph.D. student in the Edward P. Fitts Department of Industrial and Systems Engineering at North Carolina State University. He was born in Tehran, Iran. In May 2004, he graduated from Sharif University of Technology with a Bachelor of Science degree in industrial engineering. To pursue advanced degrees, he came to the United States in August 2004. He received a Master of Science degree with a dual major in industrial engineering and operations research from North Carolina State University in December 2006. While working on his Ph.D., he also served as a research assistant, instructor, and teaching assistant. His research interests include game theory, mechanism design, health economics, stochastic processes, and simulation.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisor, Dr. Stephen D. Roberts, for his kind support and thoughtful guidance through the long process of completing this dissertation. Special thanks go to Dr. Julie Ivy, Dr. Michael Pignone, Dr. Fanis Tsoulouhas and Dr. Reha Uzsoy for serving on my committee and for all the enlightening advice they have given me on this work. I also thank Dr. Marie Davidian and Dr. Jeff Joines for serving at different times as the Graduate School representative on my committee. Finally, I would like to express my sincere gratitude to my parents, my brother, and my friends for their continuous support and inspiration during my studies.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

A Game-Theoric Framework for Estimating the Willingness-To-Pay for Health and for Expansion

Abstract

A health purchaser's willingness-to-pay (WTP) for health is defined as the amount of money the health purchaser (e.g., a health maximizing public agency or a profit maximizing health insurer) is willing to spend for an additional unit of health. In this paper, we propose a game-theoretic framework for estimating a health purchaser's WTP for health in markets where the health purchaser offers a menu of medical interventions, and each individual in the population selects the intervention that maximizes her utility. We discuss how the *WTP for health* can be employed to determine the medical guidelines, and to price the new medical technologies, such that the health purchaser finds them worthwhile to implement. The framework further introduces a measure for *WTP for expansion*, defined as the amount of money the health purchaser is willing to pay for one percent of increase in the consumption level of an intervention. This measure can be employed to find how much to invest in expanding a medical technology through opening new facilities, advertising, educating the population, etc. Applying our framework to Colorectal Cancer screening tests, we estimate the WTP for health to be $\$9,950$ per quality-adjusted life years, and the WTP for expanding Colonoscopy to be \$45.40 per person per percent increase, for the 2005 U.S. population.

Key words: game theory, willingness-to-pay, health care, simulation, colorectal cancer, resource allocation, implementation.

1.1 Introduction

Society's willingness-to-pay (WTP) for health is defined as the amount of money the society is willing to spend for one unit of health. Due to its essential role in health care resource allocation and cost-effectiveness analysis, a substantial body of literature has focused on estimating this parameter. The most popular WTP estimation methods are based on surveys (for example, see King et al. (2005) and Byrne et al. (2005)). In these approaches, the WTP is assessed by different techniques: the value-of-life literature, WTP per quality-adjusted life year (QALY) ratios, and the studies of health-state values that use both utility and contingent valuation methodologies based on empirical data (Diener et al. 1998, Hirth et al. 2000).

The WTP estimates obtained from these methodologies are used to find the *socially* optimal medical resource allocations and cost-effective programs (Gold et al. 1996). Such studies, which help the decision-makers allocate the limited medical resources among a number of competing medical programs to maximize the population's health, have constituted a substantial body of health care literature. Yet, there are many medical programs that have proven to be cost-effective in these studies, but have not been properly *implemented* in the society; that is, their utilization are still suboptimal due to different barriers, such as not being included in the insurers' health plans, high co-payment, limited access to the participating medical facilities, population's unawareness about the benefits of the program, etc. For instance, cancer screening tests are considered to be cost-effective but many individuals who are at risk of developing cancer do not use these tests mainly because of high co-payments, limited access to the health providers, unawareness, etc (Swan et al. 2003, American Cancer Society 2007).

In health care systems, there are two parties that make the main contribution to the implementation of a medical program: (1) health purchaser, (2) health provider. Health purchasers are institutions that enter into contractual agreement with health providers (e.g., hospitals) in which the health provider agrees to render service to the population and be reimbursed by the health purchaser according to a prespecified contract. A health purchaser can be a government (under universal health care) or a private health insurer. A health purchaser contributes to the implementation of a medical program by including the program in the offered health plan and at an affordable co-payment. The health providers also play a major role in implementing a medical program through offering and promoting the program.

This paper attempts to address two questions in the domain of *medical implementation*:

1. Consider a medical program, such as a medical product (e.g., vaccine) or a medical guideline (e.g., recommendations for cancer screening). An important question raised by health policy makers and health providers is whether their proposed policies or products will be implemented by the health purchasers; that is, whether the programs are to be included in the health plans and be sufficiently reimbursed.

2. Now, suppose that a health provider would like to contribute to the implementation of a medical program through allocating more medical resources, advertising the program, etc. In making such decision, the health provider need to know how much her attempt will be valued (and financially supported) by the health purchaser. The importance of this question is more pronounced for preventive care, which usually requires mass implementation.

Answering these two questions relies respectively on how much the health purchaser is willing to pay for health, and how much he is willing to pay for expanding the medical programs. These two types of willingness-to-pay are defined below:

1. **WTP for health** is the amount of money the health purchaser is willing to pay to obtain one additional unit of health; we denote the WTP for health by λ, and

2. **WTP for expansion** is the amount of money the health purchaser is willing to pay for one percent of increase in the consumption level of an intervention, per person; we denote the WTP for expansion by γ.

The *WTP for health* can be used by policy makers in determining medical guidelines, and by health provider in pricing a new medical technology as such to be implemented by the health purchaser. For instance, if a new medical technology demands a WTP for health greater than λ, then it is unlikely to be implemented by the health purchaser with the exact same specifications.

The *WTP for expansion* can be used in decisions regarding expanding an existing medical program. If a health provider, such as a hospital, can increase the consumption level of a medical intervention by one percent through providing easier access, advertisement, etc., then the health purchaser is willing to pay up to γ dollars per person to the health provider.

The health purchaser's WTP for health and for expansion stems from two types of costs which in order to reduce, a health purchaser is willing to pay to health providers or other parties in the market. These are the costs that a health purchaser *implicitly* incurs, besides the direct costs of implementing a medical program:

1. **Health-related costs** include the costs (or penalties) that the health purchaser incurs due to the deterioration of the population's health. These *unobservable* penalties are imposed by two factors: (1) market regulations, which are usually enforced by the government and the health-related legislations, and (2) market competition. Health-related costs increase by the level of regulations in the health care system and the competition among the health insurers.

2. **Incentive-related costs** include the costs of providing sufficient incentives to the population for undergoing the optimal care. These costs occur when the health purchaser

benefits from the population undergoing a specific medical intervention, but the population is reluctant in using the targeted intervention. The health purchaser is therefore willing to pay for actions motivating the population to undergo the targeted interventions. Those actions may include eliminating the copayment, providing easier access, advertisement, etc.

Despite having essential implications in health care decisions, the health purchaser's WTPs cannot be easily estimated, since they are mainly affected by factors that are not *quantitatively* observable, such as the level of regulation and the competition existing in the market. In this paper, "health purchaser" (HP) means any institution that operates under a certain budget (collected from premiums, taxes, etc.) with the objective of maximizing the population's health. A health purchaser can also be a profit maximizing subject to a health constraint imposed by health regulations, competition, etc. An example of the former is the "government" under universal health care, and an example of the latter is a "private health insurer" in a competitive health insurance market.

To estimate the WTP for health, Lee and Zenios (2008) proposed a "shadow price framework," in which a *perfect agent* allocates societal health resources as such to maximize the societal health. They assume that the perfect agent is a health maximizer subject to a budget constraint, and therefore at the optimal allocation, the Lagrangian multiplier of the budget constraint, called the *shadow price*, indicates the rate at which financial resources will be exchanged for health. We use a similar shadow price framework to estimate the health purchaser's WTP for health. Lee and Zenios (2008)'s model, however, considers only one intervention, while in practice a health purchaser may offer a menu of medical interventions for a given disease.

In this paper, we propose a game-theoretic framework for estimating a health purchaser 's WTPs in a market where a number of medical programs are available for a given disease from which, each individual in the population selects the intervention that yields the highest prospect for her. In this framework, the health purchaser and the population enter into a two-move game (Gibbons 1992), where the health purchaser makes the first move in offering a contract (i.e., the set of medical interventions and their coinsurance rates) and then, having observed the contract and her health status (characterized by different risk factors, such as age, race, previous history of the disease, exposure, etc.), each individual decides which medical alternative to undergo. If an equilibrium exists, we say it *implements* a given intervention if the health purchaser is willing to cover it and a portion of the population is willing to consume the intervention. By mapping the equilibrium determined by this game to the *observed* contract and the consumption level of each intervention, we estimate the health purchaser 's WTPs for health and for expansion.

The proposed framework can be applied to markets with two central properties: (1) existence of a number of *mutually exclusive* medical alternatives for the underlying disease, and (2) a population consisting of rational individuals who have freedom in selecting the alternatives that

maximize their utilities. A set of medical alternatives is mutually exclusive if each individual chooses only one medical alternative to use at a time. An example of mutually exclusive medical interventions is a set of cancer screening alternatives (regimes), where each individual uses one screening alternative at any point of time (Note that 'use nothing' may also be considered as a medical alternative). The proposed framework also allows the population to be heterogeneous, i.e., consisting of different *risk categories*. In a heterogeneous population, individuals with the same risk of contracting the disease and the same prospective devastation caused by the disease constitute a risk category.

This paper makes several contributions. (1) It provides new insights on the *implementation* of medical programs by revisiting the concept of WTP in health care, presenting well-defined implications for policy makers and health providers in determining medical programs that are going to be implemented by the health purchasers. (2) This paper introduces the *WTP for expansion* as a new measure for *moral hazard*. The proposed framework shows how the moral hazard is affected by the health purchaser's WTP for health, and that depending on how much the health purchaser is willing to pay for health, the moral hazard can be benign (i.e., positive WTP for expansion) or malignant (i.e., negative WTP for expansion). (3) Most of the current WTP methodologies pose problems and drawbacks that are well documented in the literature (for instance, refer to Gafni and Birch (2006), Cookson (2003), Polsky (2005), and Hirth et al. (2000)). The survey participants very often do not fully understand the health consequences of the disease or the related interventions under study. In addition, the effect of moral hazard can undermine the accuracy of the WTP estimates (Polsky 2005). That is, an insured person usually tends to consume more health care services than the amount she states to be willing to pay for. This paper contributes to the WTP literature by developing an alternative approach to estimating the society's WTP in markets where the health purchaser's WTP is an adequate proxy for the society's WTP (e.g., under the presence of universal health care).

The remainder of this paper is organized as follows. We formulate the health purchaser-population game in §1.2, and develop the methodologies for finding the equilibrium and the WTPs in §1.3. We employ our proposed framework to estimate the WTPs for Colorectal Cancer screening tests in §1.4. Section 1.5 concludes the paper. All the proofs are given in the appendix.

1.2 The Health Purchaser-Population Game

A market for a set of medical interventions for a given disease generally consists of the following entities: (1) health purchaser (HP) who we assume to be a *perfect agent* maximizing the population's health subject to a budget constraint, or profit maximizer subject to a health constraint, (2) population who we assume to be (rational) prospect maximizers, and (3) health

providers who are either a profit maximizer or a cost minimizer. We assume that the HP reimburses the health provider according to a prespecified *fee-for-service* contract in which $100\beta\%$ of the health provider's service price is collected from the population and the remaining $100(1-\beta)\%$ is paid by the HP. In our framework, the physicians are assumed to be *truth-telling agents* who truthfully share the consequences of using each medical alternative with their patients. The patient and her physician are assumed to constitute a *perfect agent* who chooses the medical alternative that maximizes the patient's utility (Mooney and Ryan 1993).

1.2.1 Preliminaries

We consider a disease for which M mutually exclusive medical interventions are available. We denote each medical intervention by $j = 1, \ldots, M$, the market price of each by p_j, and the indirect cost of obtaining the intervention j by c_j. The cost c_j may include the value of patient's time, travel costs, etc. Undergoing the intervention j may also cause a health-related disutility e_j due to invasive procedure such as Colonoscopy. We denote 'using no intervention' by $j = 0$, for which $p_j = c_j = e_j = 0$. Let A denote the set of available medical interventions.

The outcomes of using a medical intervention are usually measured by its effect on the financial and the health status of the patients. To measure health, we use Quality-Adjusted Life Year (QALY), which is an aggregate variable to represent both the length of a life and the health quality of the years lived (Miyamoto et al. 1998, Pliskin et al. 1980). To each individual in the population, we assign a continuous value $\theta \in [0, 1]$, called "rank," such that $\theta\%$ of population expects higher magnitude of suffering from the disease. The details of this assignment are explained in the appendix.

Let $q_j(\theta)$ return the expected gain (or loss) in QALYs during period $[t_1, \infty]$ for an individual with rank θ who uses alternative j at time t_1 (t_1 can be the current age). The function $q_j(\cdot)$ may include losses in QALY that might occur due to the disease or normal aging. Let $v_j(\theta)$ return the expected treatment costs during period $[t_1, \infty]$ for an individual with rank θ who uses alternative j at time t_1, excluding the price of the medical intervention obtained. Gains and losses should be defined relative to a status quo condition. For preventive care, we assume that the status quo condition is perfect health with no loss in money or health due to the disease. For acute illnesses, we assume that the status quo condition is the current health status, and any improvement (deterioration) in health and financial status is considered as gain (loss). The section "Continuous Approximation" in the appendix describes how to approximate the functions $q_j(\cdot)$ and $v_j(\cdot)$ for a given population. It is reasonable to assume that functions $q_j(\cdot)$ and $v_j(\cdot)$ are monotonically increasing.

1.2.2 A Model for the Population's Preferences

Expected utility theory has been traditionally used to model the individuals' choices in selecting among a number of alternatives (Keeney and Raiffa 1993). However, in the context of health care, many empirical studies have shown that expected utility theory is not an adequate descriptive model for patients' choices (for instance, refer to Bayoumi and Redelmeier (2000) and Seror (2008)). These studies show that Prospect Theory (Kahneman and Tversky 1979, Tversky and Kahneman 1992) consistently provides a better model for the patients' choices. Following the Prospect Theory's framework, we therefore assume that an individual ranks different health profiles based on the prospective monetary loss, and gain or loss in health.

Let $\succsim$ be a preference relation on $\mathbb{R} \times \mathbb{R}^-$ representing an individual's preferences over the possible health and monetary outcomes. For an individual with rank θ, let $q_\theta \in \mathbb{R}$ denote the prospective gain (or loss) in QALY, and let $v_\theta \in \mathbb{R}^-$ denote the prospective monetary loss. We assume that the socioeconomic factors of an individual do not have any direct effect on the individual's preference over health and monetary outcomes. That is, the individual ranks different outcomes solely based on the prospective gain (or loss) in QALYs, q_θ, and the prospective monetary lost, v_θ.

Let $u(\cdot)$ be a value function on $\mathbb{R}\times\mathbb{R}^-$ that represents $\succsim$; i.e., $(q_\theta, v_\theta) \succsim (q'_\theta, v'_\theta) \Leftrightarrow u(q_\theta, v_\theta) \geq u(q'_\theta, v'_\theta)$. We assume that the preference relation $\succsim$ satisfies the necessary conditions specified by Zank (2001) so that the function $u(q, v)$ can be represented by a continuous multiplicative utility that is strictly increasing in q and v. That is, $u(q, v) = u_q(q)u_v(v)$, where $u_q(0) = u_v(0) = 1$, $u_q(q)$ is strictly increasing and convex for $q < 0$, and strictly increasing and concave for $q \geq 0$; and $u_v(v)$ is strictly increasing and convex for $v \leq 0$.

In expected utility theory, it is usually assumed and empirically shown that the marginal utility of wealth rises with better health (Klose 2003, Hammerschmidt et al. 2004). In the gain and loss environment, this condition is equivalent to assuming that the marginal disutility of cost rises with better health. It is straightforward to show that the utility function defined above satisfies this condition.

We assume that each individual decides on which medical intervention to use at time t_1 (current time) in pursuit of achieving a better financial and health status at time $t_2 > t_1$. For preventive care, the period $t_2 - t_1$ is usually larger than that of acute illnesses; for instance, people use cancer screening tests at age $t_1 = 50$ in order to prevent a cancer that may become malignant at age $t_2 = 60$, if no cancer screening test had been used. On the other hand, for acute illnesses the period $t_2 - t_1$ is on the order of days or weeks.

We suppose that when making decision at time t_1, an individual assumes that during period $[t_1, t_2]$, the underlying disease does not cause any further costs or deterioration in her health status; and the effect of her decision at time t_1 will become evident after time t_2. Therefore,

if we assume that the individual's lifetime utility is represented by the sum of each decision period's utility (Smith and Keeney 2005, Klose 2003, Hammerschmidt et al. 2004), then the utility of an individual with rank θ who uses alternative j at time t_1, can be represented by:

$$u_j(\theta) = u(-e_j, -\beta_j p_j - c_j) + U_j(\theta) = u_q(-e_j)u_v(-\beta_j p_j - c_j) + U_j(\theta), \tag{1.1}$$

where β_j is the coinsurance rate for the intervention j, and $U_j(\theta)$ represents the expected prospect at time t_2 for an individual of rank θ who uses the intervention j at time t_1. The section "Continuous Approximation" in the appendix describes how to find the functions $U_j(\theta)$ for a given population.

We assume that the population's utility is well-behaved, i.e., an individual of lower rank (i.e., with higher expected magnitude of suffering from the disease) gains more utility from using an intervention than an individual of higher rank.

Assumption 1 (well-behaved utility) *Let $C_j(\theta) = U_j(\theta) - U_0(\theta), j \in A$; the well-behaved utility assumption implies $\partial C_j/\partial\theta < 0$.*

1.2.3 A Model for the Health Purchaser's Decision

Suppose that the set of M medical interventions divides the population into $M+1$ intervals $[x_{j+1}, x_j]$, for $j = M, \ldots, 0$, and $0 = x_{M+1} < x_M < x_{M-1} < \ldots < x_1 < x_0 = 1$, such that, for $\theta \in [x_{j+1}, x_j]$, the individual of rank θ obtains the medical intervention j. Note that for $\theta \in [x_1, x_0]$, no medical intervention is used. We assume that the HP is able to manipulate the variable x_j through some mechanism (e.g. setting co-payments, advertisement, etc.). Let $V_j(x_j)$ and $Q_j(x_j)$ denote the total gain in treatment costs and the gain in quality of life, respectively, if individuals of rank $\theta \in [x_{j+1}, x_j]$ uses alternative j. Hence, the functions $V_j(\cdot)$ and $Q_j(\cdot)$ can be calculated by

$$V_j(x_j) = \int_{x_{j+1}}^{x_j} \big((1-\beta)v_j(\theta) - (1-\beta_j)p_j\big)\, d\theta, \text{ and } Q_j(x_j) = \int_{x_{j+1}}^{x_j} (q_j(\theta) - e_j)d\theta,$$

where β is the coinsurance rate for any other costs incurred due to the underlying disease except for the medical interventions in set A.

A profit maximizer HP solves the following optimization problem to determine the proportion of population for which the intervention j should be implemented:

$$\max_{0=x_{M+1}<\ldots<x_1<x_0=1} \pi = \sum\nolimits_{j=0}^{M} V_j(x_j) \text{ s.t. } \sum\nolimits_{j=0}^{M} Q_j(x_j) \geq Q^0, \tag{1.2}$$

where Q^0 is the minimum improvement in population health that the HP want to sustain while determining the value of x_j's. The *implicit* threshold Q^0 is essentially due to market competition

and the government health regulations. In the optimization problem (1.2), the health constraint is apparently binding at the optimum, since otherwise, the HP could increase his objective function by changing some of the variables x_j without violating the health constraint. Also note that in problem (1.2) if at a solution, $x_{j+1} = x_j$, implying that intervention j is not implemented, we can remove the intervention j from set A to preserve the existence of the well-behaved intervals $[x_{j+1}, x_j]$ with $x_{j+1} \neq x_j$, for all j.

At optimality of problem (1.2), the shadow price of the health constraint, denoted by $\lambda^* \geq 0$, represents the amount of increase in the objective function if the HP reduces Q^0 by one unit of health. For instance, assume that a health provider introduces a new technology which can be implemented at zero cost by the HP, and it improves the population's health by Q_{M+1} units of health. Such improvement is equivalent to reducing Q^0 by Q_{M+1} units, which results in a gain of $\lambda^* Q_{M+1}$ dollars for the HP. Hence the HP is willing to pay up to $\lambda^* Q_{M+1}$ dollars for the new technology.

In some markets, specifically in the presence of universal health care, the HP is better modeled by assuming that he is a health maximizer subject to a budget constraint. Hence, the HP solves the following optimization problem:

$$\max_{0=x_{M+1}<\ldots<x_1<x_0=1} \varpi = \sum\nolimits_{j=0}^{M} Q_j(x_j) \text{ s.t. } \sum\nolimits_{j=0}^{M} V_j(x_j) \geq B^0,$$

where B^0 is the maximum budget that the HP is willing to allocate for the set of interventions under study. In this optimization problem, the shadow price of the constraint, denoted by $\mu^* \geq 0$, represents the HP's willingness-to-take, defined as the amount of health that the HP is willing to sacrifice to relax his budget constraint by one monetary unit.

Theorem 1 *If a health purchaser is a health maximizer subject to a budget constraint or a profit maximizer subject to a health constraint, then there exists a unique constant λ^* such that the health purchaser's optimization problem at the optimal allocation $\mathbf{x}^* = (x_1^*, \ldots, x_M^*)$ will be equivalent to*

$$\max_{0=x_{M+1}<\ldots<x_1<x_0=1} L(x_1, \ldots, x_M) = \sum\nolimits_{j=0}^{M} V_j(x_j) + \lambda^* \sum\nolimits_{j=0}^{M} Q_j(x_j), \tag{1.3}$$

where $\lambda^* \geq 0$ is the health purchaser's willingness-to-pay for health. The health purchaser's willingness-to-take is represented by $\mu^* = 1/\lambda^*$.

Theorem 1 states that if we assume the *observable* allocation $\mathbf{x}^* = (x_1^*, \ldots, x_M^*)$ (i.e., the consumption level of each medical alternative in the current market) is optimum for the HP, then there exists a constant λ^* such that the optimal solution of the optimization problem (1.3) maps exactly to the observable allocation $\mathbf{x}^* = (x_1^*, \ldots, x_M^*)$; such λ^* will represent the HP's

WTP for health. The significance of this theorem lies in the fact that even though the HP's budget is not observable, or the amount of competition and regulation in the market is not quantifiable, once can still estimate the HP's WTP for health by what can be observed in the market, i.e., the consumption level of each medical alternative.

By modeling the HP's play according to Theorem 1 we implicitly assume that the HP is risk-neutral, which is justified if the medical technologies are implemented for a large number of people.

1.2.4 Formulating the Health Purchaser-Population Game

We model the interaction between the HP and the population as a two-move game:

Stage 1: The HP moves first and offers the contract $(\beta_1, \ldots, \beta_M)$, through which he specifies the interventions to be covered for the underlying disease and the corresponding co-insurance rates β_j.

Stage 2: Having observed the contract, each individual decides which medical intervention to use.

Note that Stage 1 is in fact part of the HP's health insurance plan, which consists of all medical programs being covered and their co-payments. Thus very often, Stage 1 is reached once the individual purchases a health insurance plan (for instance, from a private insurance company) or becomes eligible for a health insurance plan (for instance, offered by Medicare). On the other hand, Stage 2 is reached whenever the individual enters into a health status for which using the interventions in set A is recommended (e.g. using Colorectal cancer screening tests for individuals of age equal to or greater than 50.).

By utility function (1.1), an individual of rank θ chooses her preferred medical alternative (J_θ) according to the following optimization problem (See Figure 1.1(a) for an illustration):

$$J_\theta = \underset{j \in A+\{0\}}{\arg\max} \big(u_q(-e_j)u_v(-\beta_j p_j - c_j) + U_j(\theta)\big), \tag{1.4}$$

Let $\pi_j(\theta, \beta_j)$ denote the HP's net monetary benefit for an individual of rank θ who uses medical alternative j. By Theorem 1, if the HP's WTP for health is λ, then $\pi_j(\theta, \beta_j) = (1-\beta)v_j(\theta) - (1-\beta_j)p_j + \lambda(q_j(\theta) - e_j)$, for $j = 1, \ldots, M$, and $\pi_0(\theta) = (1-\beta)v_0(\theta) + \lambda q_0(\theta)$. For any given contract $(\beta_1, \ldots, \beta_M)$, the HP solves the optimization problem (1.4) to find the choice function $J_\theta(\beta_1, \ldots, \beta_M)$ for an individual with rank θ. Then, to find the optimal contract,

the HP maximizes his net monetary benefit by solving:

$$\max_{\beta_1,\ldots,\beta_M} \Pi = \int_{\theta\in[0,1]} \pi_{J_\theta}(\theta, \beta_{J_\theta}) d\theta \tag{1.5}$$

$$J_\theta = \arg\max_{j\in A+\{0\}} \big(u_q(-e_j)u_v(-\beta_j p_j - c_j) + U_j(\theta)\big) \text{ for } \theta \in [0,1] \tag{1.6}$$

$$0 \le \beta_j \le \beta_j^{\max}, \text{ for } j = 1,\ldots,M, \tag{1.7}$$

where $\beta_{\max}$ is the maximum coinsurance rate that the HP may declare in the context of the underlying disease (in common practice, $\beta_j^{\max}$ is usually less than 50%).

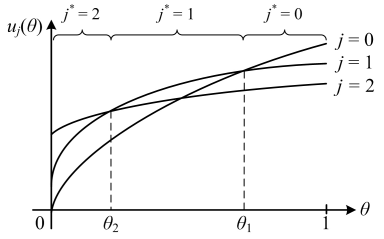
Alternatively, one may model the interaction between the HP and the population in a principal-agent framework (Laffont and Martimort 2001), in which the HP (principal) tries to design the contract $(\beta_1,\ldots,\beta_M)$ as such to induce the individual with rank θ (agent) to reveal her rank by selecting the intervention that is specifically designed for her, i.e., J_θ. This approach also leads to solving the same optimization problem as (1.5)-(1.7). Furthermore, this problem may also be related to the optimal design of insurance contracts. The studies on the insurance contract design, however, do not generally capture the properties of our problem; while we consider the design of the insurance contract in the context of a *single* disease, those studies usually investigate the design of an insurance contract from a higher level perspective, i.e., determining coverage, deductibles and premiums; likewise, the problem of moral hazard and adverse selection have also been investigated in such a high-level context (for instance, refer to Ellis and Manning (2007), Chernew et al. (2000), Blomqvist (1997) and the references therein).

1.3 Estimating the Willingness-To-Pay

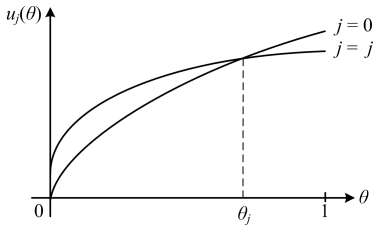
To estimate the WTPs, we first characterize the equilibrium of the HP-population game, and then by mapping the equilibrium of this game to the *observed* equilibrium, we estimate the WTPs. In §1.3.1, we assume that only one intervention is available and find closed-form expressions for the WTPs. This assumption is relaxed in §1.3.2.

1.3.1 WTPs for One Intervention

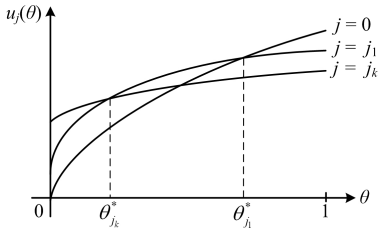
When only one intervention is available, say j, each contract (β_j) splits the population into two groups: the individuals with rank $\theta \in [0, \theta_j^*]$ who use the intervention, and the individuals with rank $\theta \in [\theta_j^*, 1]$ who do not use the intervention. The HP's problem is to find the splitting point θ^* by offering an appropriate contract as to maximize his net monetary benefit. For one intervention, the optimal contract for intervention j (β_j^*) and the optimal coverage interval



(a) Population choice function



(b) The optimal contract for intervention j



(c) The optimal contract for interventions j_1 and j_k

Figure 1.1: Optimal Contracts

$[0, \theta_j^*]$ can be determined by solving the following optimization problem (see Figure 1.1(b)):

$$\max_{\beta_j \geq 0, \theta_j \geq 0} \quad \Pi = \int_0^{\theta_j} \left(\pi_j(t, \beta_j) - \pi_0(t)\right) dt, \tag{1.8}$$

$$\text{s.t.} \quad u_q(-e_j) u_v(-\beta_j p_j - c_j) + U_j(\theta_j) = U_0(\theta_j), \tag{1.9}$$

$$\theta_j \leq 1, \tag{1.10}$$

$$\beta_j \leq \beta_j^{\max}, \tag{1.11}$$

$$\Pi \geq 0. \tag{1.12}$$

The contraint (1.9) specifies the splitting point θ_j^* (see Figure 1.1(b)), and constraint (1.12) guarantees the HP's interest in implementing the intervention. Let τ, η, and δ denote the Lagrangian multiplier of constraints (1.9), (1.10) and (1.11), respectively, and define the function $C_j(\cdot)$ as $C_j(\theta) = U_j(\theta) - U_0(\theta)$. The optimal solution (β_j^*, θ_j^*) must satisfy the following first-order conditions:

$$p_j \theta_j^* + \tau p_j u_q(-e_j) u_v'(-\beta_j^* p_j - c_j) - \delta \begin{cases} = 0, \text{ for } \beta_j^* > 0, \\ < 0, \text{ for } \beta_j^* = 0; \end{cases} \tag{1.13}$$

$$\pi_j(\theta_j^*, \beta_j^*) - \pi_0(\theta_j^*) - \tau \left.\frac{dC_j}{d\theta}\right|_{\theta = \theta_j^*} - \eta \begin{cases} = 0, \text{ for } \theta_j^* > 0, \\ < 0, \text{ for } \theta_j^* = 0; \end{cases} \tag{1.14}$$

$$\eta(\theta_j^* - 1) = 0; \tag{1.15}$$

$$\delta(\beta_j^* - \beta_j^{\max}) = 0. \tag{1.16}$$

Substituting the observed equilibrium (β_j^*, θ_j^*) into the first-order conditions (1.13)-(1.16) yields the HP's WTPs.

Theorem 2 *Given the allocation level θ_j^*, the health purchaser's WTP for expansion is represented by the Lagrangian multiplier τ multiplied by $C_j'(\theta_j^*)$. That is, $\gamma^* = \tau^* C_j'(\theta_j^*)$.*

The proposition below summarizes the WTP estimations for the cases commonly occuring in practice (i.e., cases where some portion of population uses the intervention).

Proposition 1 *Let (β_j^*, θ_j^*) denote the equilibrium of the game (1.8)-(1.12).*

1. *If at equilibrium, $0 < \beta_j^* < \beta_j^{\max}$ and $0 < \theta_j^* < 1$, then*

   (a) *the health purchaser's WTP for expansion is equal to*

   $$\gamma^* = \frac{-\theta_j^* C_j'(\theta_j^*)}{u_q(-e_j)u_v'(-\beta_j^* p_j - c_j)}, \tag{1.17}$$

   which is positive under the well-behaved utility Assumption 1; and

   (b) *the health purchaser's WTP for health is equal to*

   $$\lambda^* = \frac{\gamma^* - (1-\beta)(v_j(\theta_j^*) - v_0(\theta_j^*)) + (1-\beta_j^*)p_j}{q_j(\theta_j^*) - e_j - q_0(\theta_j^*)}. \tag{1.18}$$

2. *If at equilibrium, $\beta_j^* = 0$ and $0 < \theta_j^* < 1$, then the health purchaser's WTP for expansion is at least equal to γ^* in Eq. 1.17, and the health purchaser's WTP for health is at least equal to the λ^* in Eq. 1.18.*

3. *If at equilibrium, $\beta_j^* = \beta_j^{\max}$ and $0 < \theta_j^* < 1$, the health purchaser's WTP for expansion is at most equal to the γ^* in Eq. 1.17 and the health purchaser's WTP for health satisfies Eq. 1.18 and $\Pi^* \geq 0$.*

In Proposition 1, cases (1) and (2) are referred to as *benign* moral hazard (Pauly and Held 1990), where the HP's WTP for expansion is positive. That is, the HP benefits from expanding the program to a larger portion of population. In case (3), the sign of the HP's WTP for expansion can be either negative or positive. If we assume that the population's utility is well-behaved (Assumption 1) then by expression (1.14), the sign of the HP's WTP for expansion is the same as the sign of $\pi_j(\theta_j^*, \beta_j^*) - \pi_0(\theta_j^*)$ (recall that $0 < \theta_j^* < 1$). Therefore, if the HP finds expanding the program *marginally* cost-effective, i.e., $\pi_j(\theta_j^*, \beta_j^*) - \pi_0(\theta_j^*) > 0$, then his WTP

for expansion will be positive (benign moral hazard), whereas if $\pi_j(\theta_j^*, \beta_j^*) - \pi_0(\theta_j^*) < 0$, the HP's WTP for expansion will be negative, implying that attempts for expanding the program will not be supported by the HP.

When $\gamma^* > 0$, the HP is willing to pay up to γ^* per person for a percent of increase in the consumption level of the intervention j. For instance, if an action increases the consumption of intervention j from θ_j^* to $\theta_j^* + \epsilon$, the HP is willing to pay up to $\epsilon\gamma_j^*$ per individual in the population for such an action.

1.3.2 WTPs for Multiple Interventions

When more than one intervention is offered, the equilibrium of the HP-population game cannot be characterized by a closed-form solution. In this section, we first develop a methodology for finding the equilibrium in general case and then we show how to estimate the WTPs.

Optimal Allocation Algorithm

In general case of M interventions, the equilibrium of the HP-population game can be found by solving the daunting optimization problem (1.4)-(1.7). The following proposition allows us to develop an efficient algorithm for solving this problem.

Proposition 2 *In the sequence of interventions $s = (j_1, \ldots, j_{k-1}, j_k, j_{k+1}, \ldots, j_M)$, if for the first $k-1$ interventions, the optimal coinsurance rates, i.e., $(\beta_{j_1}^*, \ldots, \beta_{j_{k-1}}^*)$, and the optimal allocation levels, i.e., $(\theta_{j_1}^*, \ldots, \theta_{j_{k-1}}^*)$, are known, then solving the following optimization problem yields the optimal coinsurance rate $(\beta_{j_k}^*)$ and the optimal allocation level $(\theta_{j_k}^*)$ of the intervention j_k:*

$$\max_{\beta_{j_k}, \theta_{j_k}} \quad \Pi_{j_k} = \int_0^{\theta_{j_k}} \left(\pi_{j_k}(t, \beta_{j_k}) - \pi_{j_{k-1}}(t, \beta_{j_{k-1}}^*) \right) dt, \tag{1.19}$$

$$s.t. \quad u(-e_{j_k}, -\beta_{j_k} p_{j_k} - c_{j_k}) + U_{j_k}(\theta_{j_k}) = u(-e_{j_{k-1}}, -\beta_{j_{k-1}}^* p_{j_{k-1}}) + U_{j_{k-1}}(\theta_{j_k}), \tag{1.20}$$

$$0 \le \theta_{j_k} \le \theta_{j_{k-1}}^*, \tag{1.21}$$

$$0 \le \beta_{j_k} \le \beta_{j_k}^{\max}, \tag{1.22}$$

$$\Pi_{j_k} \ge 0. \tag{1.23}$$

If the problem (1.19)-(1.23) results in $\theta_{j_k}^* = \theta_{j_{k-1}}^*$, to reach optimality, the intervention j_{k-1} should be dropped from the sequence s.

To find the equilibrium, we first generate all possible permutations of the available interventions, since as we will see later, each permutation results in different allocation and net

monetary benefit for the HP. Each permutation, denoted by s, is called a sequence of interventions. The HP selects to implement the sequence that maximizes his NMB. Let S denote the set of all possible sequences of interventions. We select a sequence $s \in S$. At stage 1, we pick the first intervention in the sequence, denoted by j_1, and find the optimal contract $\beta^*_{j_1}$ and the optimal coverage interval $[0, \theta^*_{j_1}]$ by solving the optimization problem (1.8)-(1.12), assuming that no other intervention exists (see Figure 1.1(b)). Having fixed $\beta^*_{j_1}$ and $[0, \theta^*_{j_1}]$, we proceed to the next stage. In stage $k > 1$, we select the next intervention in sequence s, denoted by j_k, and solve the optimization problem (1.19)-(1.23) for the intervention j_k, selected in this stage, and the intervention j_{k-1}, fixed at the previous stage, to find the optimal contract for the intervention j_k, β^*_k, and the optimal coverage interval $[\theta^*_{j_k}, \theta^*_{j_{k-1}}]$ (See Figure 1.1(c)):

We continue until the optimal contracts and the coverage intervals for all interventions in the current sequence are determined. Next, we select a different sequence from the set S and repeat the procedure. The optimal sequence is the one that maximizes the HP's net monetary benefit. At each stage k, solving the problem (1.19)-(1.23) results in one of the following cases:

1. If the problem does not have a feasible solution, we remove the intervention j_k from the current sequence s, and go to the next stage with a new intervention from s assigned to j_k, and j_{k-1} remaining unchanged.

2. If $0 < \theta^*_{j_k} < \theta^*_{j_{k-1}}$, we have an interior optimal solution; we record the coinsurance rate of the test j_k and its coverage interval, and go to the next stage.

3. If $\theta^*_{j_k} = \theta^*_{j_{k-1}}$, the intervention j_{k-1} is entirely dominated by intervention j_k and j_{k-2}. We remove the intervention j_{k-1} from the current sequence and solve the optimization problem (1.19)-(1.23) for the interventions j_k and j_{k-2}.

4. If $\theta^*_{j_k} = 0$, we stop at this stage, since there is no portion of the population left to assign a new intervention to. We record the resulting HP's net monetary benefit, and proceed with a new sequence from set S.

This procedure, referred to as Algorithm 1, is outlined in Table 1.1. In Algorithm 1, as the number of interventions (M) increases, the number of treatment sequences to be evaluated increases in the order of factorial ($M!$). Yet, two properties of Algorithm 1 improve its efficiency in finding the equilibrium: (1) this algorithm converts the original problem, which has $2 \times M$ variables (β_j and θ_j, for $j = 1, \ldots, M$), into M *one-variable* optimization problems which can be efficiently solved; (2) Step 2.(c) (see Table 1.1) very often results in $\theta^*_{j_k} = 0$, which consequently terminates step 2, without investigating all the intervention in the current sequence. In the numerical application presented in §1.4, where $M = 3$, the algorithm finds the optimal solution with three-decimal accuracy in approximately 0.1 second.

Table 1.1: Algorithm 1

Step 1 : Generate all possible sequences of interventions and put them in set S.

Step 2 : `For each sequence` $s \in S$ do:

1. Set $k = 1$, `toStop = False`, $\beta^*_{j_{k-1}} = 0$, $\theta^*_{j_{k-1}} = 1$, and $\Pi^*_{j_{k-1}} = 0$.
2. Do While `toStop = False`:
   (a) Assign a new (next) intervention in sequence s to j_k.
   (b) Solve the optimization problem (1.19)-(1.23) for j_k and j_{k-1} to find $\theta^*_{j_k}$, $\beta^*_{j_k}$ and $\Pi^*_{j_k}$.
   (c) If no feasible solution is available, remove intervention j_k from sequence s, set $\Pi^*_{j_k} = 0$, and go to the next stage with a new intervention form s assigned to j_k.
   Else if $\theta^*_{j_k} = \theta^*_{j_{k-1}}$, remove intervention j_{k-1} from sequence s, set $j_{k-1} = j_{k-2}$, and restart from step (b).
   Else if $\theta^*_{j_k} = 0$, remove the remaining interventions from sequence s, and set `toStop = True`.
   (d) Update $\Pi^*_s = \Pi^*_{j_k} + \Pi^*_s$.
3. Record Π^*_s and the vectors $(\theta^*_{j_k}, \beta^*_{j_k})$ for each j_k in sequence s.

Step 3 : Select the sequence in set S with maximum Π^*_s.

Estimating the WTPs

For $M = 1$, the WTPs can be estimated by the first-order condition (1.13)-(1.16), as also summarized in Proposition 1. However, for $M > 1$, no closed-form solutions exist for calculating the WTPs. For these cases, we use the following approach to find the WTPs.

Let A^* denote the set of medical interventions that are implemented in the equilibriums of the HP-population game found by Algorithm 1, and let $\hat{A}^*$ denote the set of medical interventions that are currently implemented in the society. Let $\Lambda \subset \mathbb{R}$ represent a set that for any $\lambda \in \Lambda$, $A^* = \hat{A}^*$; that is, the λ that results in an equilibrium implementing the same set of medical interventions as those being implemented in the society (set $\hat{A}^*$). To determine the HP's WTP for health, we find the $\lambda \in \Lambda$ that results in the contract $(\beta_1, \ldots, \beta_M)$ and the allocation $(\theta_1, \ldots, \theta_M)$ which are as close as possible to the corresponding *observed* values, $(\hat{\beta}_1, \ldots, \hat{\beta}_M)$ and $(\hat{\theta}_1, \ldots, \hat{\theta}_M)$. This leads to solving the following problem:

$$\min_{\lambda \in \Lambda} \Gamma(\lambda) = \sum_{j=1}^{M} \Big(w_j(\beta_j(\lambda) - \hat{\beta}_j)^2 + z_j(\theta_j(\lambda) - \hat{\theta}_j)^2 \Big), \tag{1.24}$$

where w_j and z_j are the weights of the square errors for intervention j. Assuming that the contribution of each intervention to the HP's WTP for health is proportional to its price, it is reasonable to set $w_j = z_j = p_j$.

Having found the WTP for health from problem (1.24), one can use the following proposition to calculate the WTP for expansion for each intervention.

Proposition 3 *Suppose that at λ^* WTP for health, the sequence of intervention $(j_1, \ldots, j_M)$ is implemented with allocation levels $(\theta^*_{j_1}, \ldots, \theta^*_{j_M})$. The health purchaser's WTP for expansion for intervention j, γ^*_j , is represented by $\gamma^*_j = \tau^*_j \big(U'_{j_k}(\theta^*_{j_k}) - U'_{j_{k-1}}(\theta^*_{j_k})\big)$, where τ^*_j is the Lagrangian multiplier of constraint (1.20).*

Given the arguments in the paper, it is immediate that two health purchasers have the same WTP for health if (1) their populations have the same *texture*, i.e., the same proportion (not size) of the population has the health profile $\phi_i \in \Phi$ (refer to EC.1 for the definition of health profile), and (2) the health purchasers offer the same set of medical interventions with the same coinsurance rates for the underlying disease.

1.4 An Illustrative Application

In this section, we employ the proposed game-theoretic framework to estimate the WTPs for Colorectal Cancer (CRC) screening tests. We first provide a brief introduction to Colorectal Cancer and then present the results.

1.4.1 Colorectal Cancer

Colorectal Cancer (CRC) - cancer of the colon or rectum - is the second leading cause of cancer-related deaths in the U.S. (Centers for Disease Control and Prevention 2008). Like many other cancers, CRC becomes symptomatic when the cancer is at an advanced level and consequently the chance of survival is significantly low. However, the CRC screening tests can identify the cancer at an earlier stage leading to improved survival and considerably lower treatment cost. Common screening tests include Fecal Occult Blood Test (FOBT), Flexible Sigmoidoscopy, and Colonoscopy. The CRC screening tests have different performance in terms of cost and effectiveness, which are well examined in the literature (for instance, refer to Pignone et al. (2002) and Levin et al. (2008)).

It is generally recommended to begin screening for CRC soon after turning 50 and continue getting screened at regular intervals (Levin et al. 2008). However, as recent studies show, there are many people who are at risk for CRC and still not screened (Swan et al. 2003, American Cancer Society 2007). The patient preference for CRC screening is strongly sensitive to out-of-pocket costs, including co-payments and the indirect costs of obtaining the screening tests (Pignone et al. 1999). Consequently, among the factors influencing the tendency of people to use screening tests, the role of health insurance and accessibility to screening facilities has gained considerable attention in the recent studies. Evidence suggests that addressing insurance and cost-related barriers to care is a critical component of the effort to improve the cancer prevention and early detection practice (Ward et al. 2008, Harvard Center for Cancer Prevention 2008).

We applied our proposed game-theoretic framework to the market for CRC screening tests to answer three main questions: given the current consumption levels and coinsurance rates for CRC screening tests (1) what is the WTP for health for CRC screening tests? (2) how much is the health insurance system willing to pay for expanding the CRC screening facilities? and (3) how do different regulations affect the market of CRC screening tests?

In this section, we refer to the "health purchaser" as "health insurance system," since in the U.S., the CRC screening tests are consumed by individuals covered by different health insurers (including, private health insurers and government-sponsored health insurers), who usually offer the same set of CRC screening tests with common coinsurance rates.

1.4.2 Vanderbilt/NC State CRC Model

To find the equilibrium of the HP-population game for CRC screening tests, we obtain the necessary data from a medical simulation model called Vanderbilt/NC State model, which is a stochastic, discrete-event simulation model of the natural history of CRC (Roberts et al. 2007). This model simulates a population over time which may include a mixture of patients with different birth years, races, genders, and family histories of CRC. Screening can intervene in

the CRC process by detecting adenomas and early cancers, changing the future outcomes. The model produces discounted costs and QALYs for screening decisions as the primary outcomes and has been used to determine the cost-effectiveness of CRC screening alternatives (Tafazzoli et al. 2009).

For this paper, the screening regimens are updated according to the 2008 joint guideline from American Cancer Society, US Multi-Society Task Force on Colorectal Cancer, and the American College of Radiology (Levin et al. 2008). The details of these screening regimens are provided in the appendix.

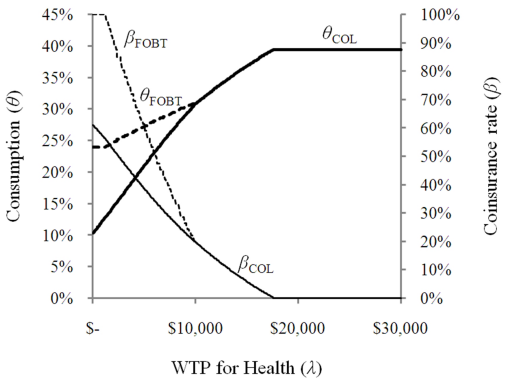
1.4.3 The Market Specifications for the CRC Screening Tests

To constitute our population, we characterize a health profile by four risk factors: age, family history, race, and gender. Four age groups, 40 to 70 incremented by 10, and two races, white and black, were considered. Hence, the defined population consists of 32 different health profiles. The proportions of the population in each health profile, α_i, were determined according to the U.S Census 2000. We consider three screening tests available in the market: FOBT, Sigmoidoscopy, and Colonoscopy, which are assumed to cost \$4.54, \$220, and \$661, respectively (Roberts et al. 2007). The indirect cost of obtaining FOBT, Sigmoidoscopy, and Colonoscopy are assumed to be \$10, \$150, and \$432 (Jonas et al. 2008). For CRC screening alternatives, the QALY functions, $q_j(\cdot)$, and the treatment cost functions, $v_j(\cdot)$, were approximated by exponential functions of form $f(x) = a + be^{cx}$. The details are provided in the appendix.

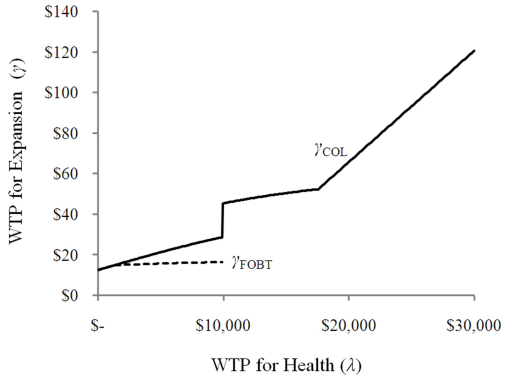
We assume that the preference relation $\succsim$ defined in §1.2.2 satisfies the necessary conditions for the value function $u(q, v)$ to be represented by a multiplicative exponential utility (Zank 2001). That is, $u(q, v) = (e^{r_q q})(e^{r_v v}) = e^{r_q q + r_v v}, q \leq 0, v \leq 0$, where $r_q > 0$ and $r_v > 0$ represent the population's health and monetary risk aversion, respectively. For simplicity, we assumed that individuals treat the probabilities linearly in the prospects. We chose $r_q = 2.75$ and $r_v = 2.75 \times 10^{-4}$, for two reasons: (1) The error of mapping the equilibrium of HP-population game to the observed equilibrium (represented by the objective function of problem (1.24)) is minimized when $r_q = 2.75$ and $r_v = 2.75 \times 10^{-4}$; i.e., these settings yield the closest equilibrium to the current consumption of CRC screening tests and the coinsurance rates. (2) In our population, setting $r_q = 2.75$ and $r_v = 2.75 \times 10^{-4}$ results in average WTP of \$28 for FOBT and \$48 for Sigmoidoscopy[1]. These numbers are in line with findings of Frew et al. (2001) in which the average population's WTP for FOBT and Sigmoidoscopy is estimated to be \$30 − \$50, independent of the screening protocol[2]. A sensitivity analysis on the values of

[1]Remember that for an individual of rank θ, the WTP for a given intervention j, denoted by WTP_j, solves $u(-e_j, -WTP_j) + U_j(\theta) = U_0(\theta)$ (See Eq. 1.1).

[2]Remember that "WTP for health" and "WTP for a medical intervention" are totally distinct concepts. The WTP for health is defined as the amount of money that the decision maker (here HP) is willing to spend for one unit of health, while the WTP for a medical intervention is the amount of money that the patient is willing to



(a) Equilibrium



(b) WTP for expansion vs. WTP for health

Figure 1.2: Equilibrium and WTP for health and expansion for $\beta_j \in [0, 100\%]$

parameters r_q and r_v is provided in the appendix.

1.4.4 Results and Analyses

Figure 1.2(a) depicts the equilibriums for different values of the WTP for health (λ), assuming that the insurers can select the coinsurance rates to be of any value in the range $[0, 100\%]$. When $\lambda = 0$, the insurer sets the coinsurance rates for FOBT and Colonoscopy at 100% and 61%, respectively. At these rates, 14% of population uses FOBT and 10% of population uses Colonoscopy. As λ increases, coinsurance rates for both tests drops while their consumption levels increase. At $\lambda = \$9,920$, the insurer finds offering FOBT no longer beneficial and offers a contract that implements only Colonoscopy. Note that according to Figure 1.2(a), Sigmoidoscopy is never implemented for any $\lambda \in [0, \$30,000]$. This finding is in line with a number of previous studies predicting that Sigmoidoscopy is unlikely to be successfully implemented and adequately reimbursed by health insurers (Lewis and Asch 1999, Shaheen and Ransohoff. 1999).

In the current market of CRC screening tests, the explanatory power of FOBT in estimating the WTP for health is not significant: first, it is a very inexpensive screening test (costs less than $10), and second, recent statistics show that the consumption rate for FOBT is declining (National Cancer Institute 2007) and Colonoscopy tends to become the preferred choice of CRC screening. Therefore, we use only Colonoscopy for estimating the WTP for health. For our population in year 2005, using the data provided by Smith et al. (2008), we estimate the percent of insured adults of ages 40 and greater who have had Colonoscopy within the last 10 years to

pay to remain on the same utility curve as not using the intervention at all (refer to footnote 1). While the WTP for health is the direct result of limited medical resources, the WTP for a medical intervention is the result of the patient's preference over monetary and health-related outcomes.

be 30% (a sensitivity analysis on this parameter is provided in the appendix). The coinsurance rate for Colonoscopy is commonly 20% (Centers for Disease Control and Prevention 2010). By following the procedure described in §1.3.2, the WTP for health for CRC screening tests is estimated to be $\lambda^* = \$9,950$ per QALY.

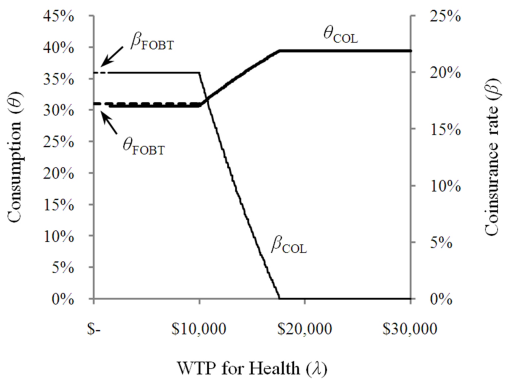
Figure 1.2(b) displays the WTP for expansion (γ) versus WTP for health (λ). As λ increases, γ also increases for both tests but at different rates. When λ is low ($< \$1,420$), the insurer system is willing to pay more for expanding FOBT than Colonoscopy. At $\lambda = \$9,920$, FOBT leaves the market and the WTP for expanding FOBT will be captured by Colonoscopy and therefore, we see a jump in the WTP for expanding Colonoscopy at this point. At $\lambda^* = \$9,950$, the WTP for expanding Colonoscopy is $\gamma^*_{\text{COL}} = \$45.40$ per percent per person, which means that the insurance system is willing to pay up to \$45.40 per individual in the population for any action that increases the consumption level of Colonoscopy by one percent, such improving access by opening a new Colonoscopy suite.

To illustrate how the proposed game-theoretic framework can be employed by health policy makers in setting the market regulations, we consider 3 markets with different specifications:

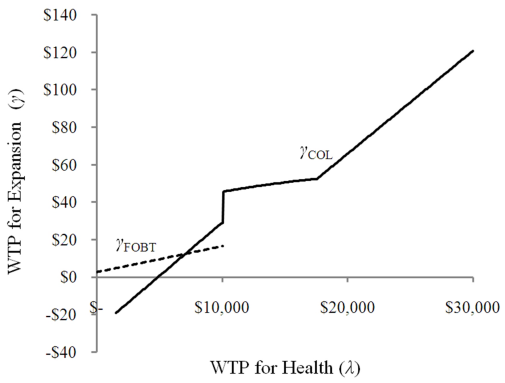
Case 1: Consider a health insurance market where the insurers are not allowed to set the coinsurance rates for FOBT and Colonoscopy exceeding 20%. Figure 1.3(a) depicts the equilibrium for different values of the WTP for health (λ). For $\lambda < \$1,540$, only FOBT is offered, with $\beta_{\text{FOBT}} = 20\%$. At $\lambda > \$1,540$, the insurers begin offering Colonoscopy. However, when $\lambda \leq \$4,920$ (see Figure 1.3(b)), the WTP for expanding colonoscopy is negative, implying that attempts for expanding the Colonoscopy consumption will not be supported by the insurance system. Therefore, reducing the coinsurance rates may not necessarily result in higher Colonoscopy utilization if the WTP for health, which is indirectly affected by the public awareness regarding the CRC screening, market competition, health regulations, etc., is not high enough.

Case 2: Now suppose that following the recommendations of Harvard Center for Cancer Prevention (2008) to eliminate the co-payments for preventive screening tests, the health insurance system is forced to set the coinsurance rates for FOBT and Colonoscopy to zero. Figure 1.4(a) depicts the equilibrium for this highly regulated market. If the WTP for health is less than $\$3,650$, the insurers will not offer Colonoscopy. The important observation in this case is on the value of WTP for expansion. According to Figure 1.4(b), for $\lambda \leq \$7,950$, although Colonoscopy is offered, the WTP for expansion is negative, implying that for $\lambda \leq \$7,950$, efforts on expanding Colonoscopy will not be supported and rewarded by the health insurance system.

Case 3: The WTP for expansion is important in determining the level of effort invested in promoting a screening test through educations, advertisement, etc., which in general increases the population's health risk aversion (r_q). Figure 1.5 shows the effect of the population's health risk aversion on the level of consumption, coinsurance rate, and the WTP for expanding the

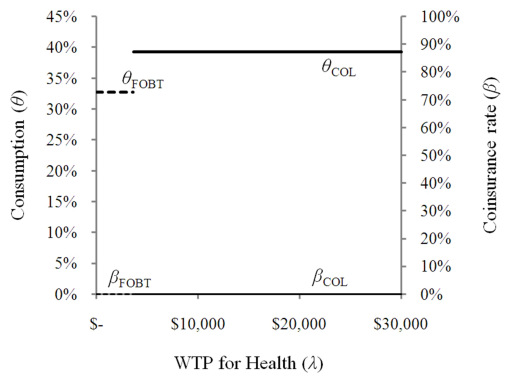


(a) Equilibrium

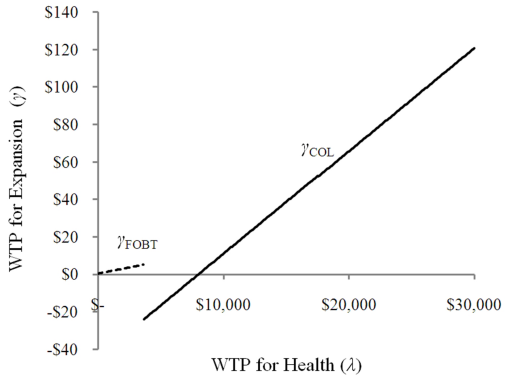


(b) WTP for expansion vs. WTP for health

Figure 1.3: Equilibrium and WTP for health and expansion for $\beta_j \in [0, 20\%]$



(a) Equilibrium

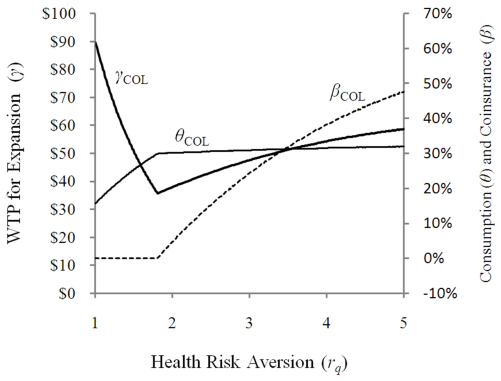


(b) WTP for expansion vs. WTP for health

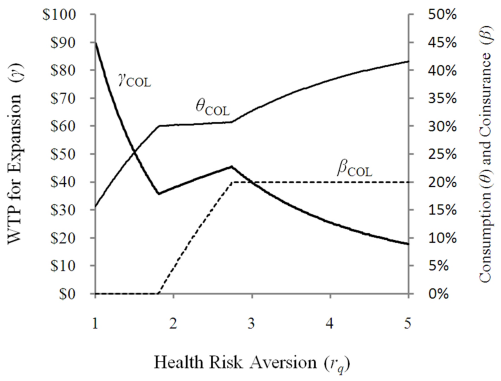
Figure 1.4: Equilibrium and WTP for health and expansion when $\beta^{\max}_{\text{FOBT}} = \beta^{\max}_{\text{COL}} = 0$

Colonoscopy. In both Figures 1.5(a) and 1.5(b), we set $\lambda = \$9,950$ and $r_v = 2.75 \times 10^{-4}$. Figure 1.5(a) depicts the equilibrium for the case where the insurers can set the coinsurance rates to any value in range $[0, 100\%]$, whereas in Figure 1.5(b) the coinsurance rate for colonoscopy cannot exceed 20%.

In Figures 1.5(a) and 1.5(b), low r_q, $1 \leq r_q \leq 2$, represents a population not sufficiently educated about the screening tests. In both figures, the WTP for expansion decreases as the population becomes more inclined to consume Colonoscopy and as a greater portion of population uses Colonoscopy. In Figure 1.5(a), since the insurers do not have any limit on setting β_{COL}, inducing higher Colonoscopy demand makes the insurers counteract the effect by increasing β_{COL} in order to keep the Colonoscopy consumption around 30%. On the other hand, in Figure 1.5(b), since β_{COL} is bounded by 20%, educating the population is in fact effective



(a) $\beta_{\text{COL}} \in [0, 100\%]$



(b) $\beta_{\text{COL}} \in [0, 20\%]$

Figure 1.5: WTP for expansion (γ)

in increasing the consumption of Colonoscopy; the WTP for expansion, however, decreases as more people get screened.

1.5 Conclusions and Future Research

We proposed a game-theoretic framework by which two types of WTP can be defined and measured in a health care system where a number of mutually exclusive medical interventions are available for a given disease. The *WTP for health* can be employed by policy makers in determining the medical guidelines, and by health providers in introducing new medical technologies, as such to be *implementable* by the health purchasers; that is, being included in their offered health plans, reimbursed and promoted. The *WTP for expansion* can be used by health providers in finding how much to invest in expanding a medical program by opening new facilities, advertising, educating the population, etc.

The proposed framework presents new insights on the implementation of medical programs: in the context of an underlying disease (1) given the estimate λ^* for the WTP for health, any new medical intervention or proposed medical guideline should demand less WTP for health than λ^* to be implemented by the health purchasers in the system, and (2) given the estimate γ^* for the WTP for expanding an intervention, the health purchaser values the contribution of a health provider to implementation of the intervention as much as γ^* monetary units per person in the population per percent of increase in the consumption level of the intervention.

In the presence of a universal health insurer or a highly regulated market, where the health purchaser's WTP for health is an adequate proxy for the society's WTP for health, our framework estimates the *society*'s WTPs for health and for expansion in the context of the underlying disease. The WTP for health defined in our framework deviates from the commonly-used WTP

methodologies, which are generally based on expected utility theory (Klose 2003) or statistical methods (Hirth et al. 2000) for deriving the society's WTP for health, but has the same shadow-price basis as the WTP for health defined by Lee and Zenios (2008). For cases where the health purchaser's WTP for health is an adequate proxy for the society's WTP for health, a study that fairs the estimate provided by our framework with other WTP estimation methods will be of great interest for future research.

One of the main limitations of our framework is the assumption of identical value functions for all health profiles (or the individual's rank). Relaxing this assumption introduces more constraints in the health purchaser-population game, making it more challenging to solve. We leave relaxing this assumption for future research. Moreover, we assumed that physicians are *truthful* agents in population; the validity of this assumption depends in part on the reimbursement contract offered by the health purchaser to the physicians. Also, as mentioned in §1.1, the framework assumes that individuals have the opportunity of deciding (perhaps through consultation with their physicians) about which medical intervention to use; therefore, the proposed framework cannot be used for cases in which the population is deprived of such opportunity (e.g. intensive care).

For the 2005 U.S. population, our framework estimates WTP for health for CRC screening tests to be $\$9,950$ per QALY, implying that any new intervention or revised guideline should demand less WTP to be implemented. The WTP for expanding Colonoscopy is estimated to be \$45.40 per person for any action (such as opening a new Colonoscopy suite, educating the population, etc.) that increases the Colonoscopy consumption by one percent. Our results show that Sigmoidoscopy is expected to leave the CRC screening test markets, as also predicted by previous studies.

The framework shows that the CRC screening utilizations can be improved by (1) increasing the WTP for health through imposing regulations (e.g., limiting the copayments and passing laws on offering the coverage of screening tests) and increasing the indirect penalties for the health purchasers (e.g., increasing the public awareness about the CRC screening tests); and (2) by increasing the population's utility from consuming screening tests through education, easier access to screening facilities, etc. The central point, however, is that these two strategies should accompany each other; if the demand for screening tests increases in a non-regulated market, the health purchasers will increase the coinsurance rates to keep the consumptions at the level optimal for them. Also, when WTP for health is not sufficiently high, decreasing or eliminating the copayments may lead to negative WTP for expansion for some screening tests, which may incent the health purchasers to stop covering or promoting the tests.

Appendices

A.1. Proof of Theorem 1

To prove Theorem 1, we need the following lemma.

Lemma 1 *Let*

$$\mathbf{A}_{k+1} = \begin{pmatrix} 0 & a_1 & a_2 & \cdots & a_k \\ a_1 & b_1 & 0 & \cdots & 0 \\ a_2 & 0 & b_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_k & 0 & 0 & \cdots & b_k \end{pmatrix}$$

then

$$|\mathbf{A}_{k+1}| = -\sum_{i=1}^{k} a_i^2 \prod_{j \neq i} b_j. \tag{1.25}$$

Proof. Suppose expression (1.25) is true for $k+1$; we show that it is also true for $k+2$.

$$|\mathbf{A}_{k+2}| = \begin{vmatrix} 0 & a_1 & \cdots & a_k & a_{k+1} \\ a_1 & b_1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_k & 0 & \cdots & b_k & \vdots \\ a_{k+1} & 0 & \cdots & \cdots & b_{k+1} \end{vmatrix} = (-1)^{k+1} a_{k+1} \begin{vmatrix} a_1 & \cdots & a_k & a_{k+1} \\ b_1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & b_k & 0 \end{vmatrix} - b_{k+1} |\mathbf{A}_{k+1}|$$

$$= (-1)^{k+1}(-1)^k a_{k+1}^2 b_1 \cdots b_k - b_{k+1}|\mathbf{A}_{k+1}| = -a_{k+1}^2 b_1 \cdots b_k - \sum_{i=1}^{k} a_i^2 \prod_{j \neq i} b_j$$

$$= -\sum_{i=1}^{k+1} a_i^2 \prod_{j \neq i} b_j. \quad \square$$

The Lagrangian function of optimization problem (1.2) is:

$$L(x_1, \ldots, x_M) = \sum_{j=1}^{M} V_j(x_j) + \lambda \left(\sum_{j=1}^{M} Q_j(x_j) - Q^0 \right). \tag{1.26}$$

The first-order optimality condition will be:

$$v_j(x_j) - v_{j-1}(x_j) + p_{j-1} - p_j + \lambda(q_j(x_j) - q_{j-1}(x_j) + e_{j-1} - e_j) = 0. \tag{1.27}$$

For a given λ, define the HP's net monetary benefit for intervention j as $\chi_j(x_j) = v_j(x_j) - p_j + \lambda(q_j(x_j) - e_j)$. Therefore, the first-order optimality condition (1.27) becomes equivalent to $\chi_j(x_j) - \chi_{j-1}(x_j) = 0$, which implies that the functions $\chi_j(\cdot)$ and $\chi_{j-1}(\cdot)$ intersect at point x_j. Also note that at optimum, if intervention j precedes intervention $j-1$, then we should have $\chi_j(x) \geq \chi_{j-1}(x)$ for $x \leq x_j$, and $\chi_j(x) \leq \chi_{j-1}(x)$ for $x \geq x_j$, since otherwise switching the place of interventions j and $j-1$ increases the objective function (1.26). Consequently, we

will have $\chi'_j(x_j) - \chi'_{j-1}(x_j) \leq 0$; i.e. at point $x = x_j$ the derivative of $\chi_j(\cdot)$ is always less than that of $\chi_{j-1}(\cdot)$.

Since at optimality, the problem (1.2) has binding constraint, the bordered Hessian matrix of Lagrangian function (1.26) is equal to:

$$\mathbf{H} = \begin{pmatrix} 0 & Q'_1(x_1) & \cdots & Q'_M(x_M) \\ Q'_1(x_1) & \chi'_1(x_1) - \chi'_0(x_1) & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ Q'_M(x_M) & 0 & \cdots & \chi'_M(x_M) - \chi'_{M-1}(x_M) \end{pmatrix}.$$

By Lemma 1 and setting $a_j = Q'_j(.)$ and $b_j = \chi'_j(x_j) - \chi'_{j-1}(x_j) \leq 0$, for $j = 1, \ldots, M$, it is immediate that the last M leading principal minor of matrix $\mathbf{H}$ alternate in sign, with the sign of the determinate of $\mathbf{H}$ matrix the same as the sign of $(-1)^M$. Therefore, the point $(x_1^*, \ldots, x_M^*)$ that satisfies the first-order Kuhn-Tucker conditions is the strict maximum (Simon and Blume 1994, Theorem 19.6 and the discussion thereafter). Hence, there exists a unique Lagrangian multiplier λ^* that makes the optimization problem (1.3) equivalent to problem (1.2) at the observed solution $(x_1^*, \ldots, x_M^*)$. That is, assuming that the observed allocation $(x_1^*, \ldots, x_M^*)$ is optimal for the HP, we can find a unique multiplier λ^* such that solving the problem (1.3) results in the exact same allocation as $(x_1^*, \ldots, x_M^*)$.

Likewise, it can be shown that when the health purchaser is health maximizer subject to a budget constraint, his optimization problem is equivalent to:

$$\max_{0=x_{M+1}<\ldots<x_1<x_0=1} K(x_1, \ldots, x_M) = \sum\nolimits_{j=0}^{M} Q_j(x_j) + \mu^* \sum\nolimits_{j=0}^{M} V_j(x_j), \tag{1.28}$$

By writing the first-order optimality conditions of problems (1.3) and (1.28), it is straight forward to show that $\mu^* = 1/\lambda^*$. The proof is then completed by multiplying the objective function (1.28) by the constant $1/\mu^*$. □

A.2. Proof of Theorem 2

We write the equality constraint (1.9) as two inequality constraints:

$$U_j(\theta_j) - U_0(\theta_j) \geq -u_q(-e_j)u_v(-\beta_j p_j - c_j), \tag{1.29}$$

$$U_j(\theta_j) - U_0(\theta_j) \leq -u_q(-e_j)u_v(-\beta_j p_j - c_j). \tag{1.30}$$

Let τ_1 and τ_2 denote the Lagrangian multiplier of constraints (1.29) and (1.30). The first-

order conditions for the new optimization problem will be:

$$p_j\theta_j^* + (\tau_2 - \tau_1)p_j u_q(-e_j)u_v'(-\beta_j^* p_j - c_j) - \delta \begin{cases} = 0, \text{ for } \beta_j^* > 0, \\ < 0, \text{ for } \beta_j^* = 0. \end{cases} \tag{1.31}$$

$$\pi_j(\theta_j^*, \beta_j^*) - \pi_0(\theta_j^*) - (\tau_2 - \tau_1)\left.\frac{dC}{d\theta}\right|_{\theta=\theta_j^*} - \eta \begin{cases} = 0, \text{ for } \theta_j^* > 0, \\ < 0, \text{ for } \theta_j^* = 0. \end{cases} \tag{1.32}$$

$$\eta(\theta_j^* - 1) = 0 \tag{1.33}$$

$$\delta(\beta_j^* - \beta_j^{\max}) = 0 \tag{1.34}$$

Comparing the first order condition (1.31)-(1.34) with (1.13)-(1.16) reveals that at optimality $\tau^* = \tau_2^* - \tau_1^*$. First we show that $-\tau^* > 0$ implies a positive WTP for expansion and $-\tau^* < 0$ implies a negative WTP for expansion. Suppose $-\tau^* > 0$. Therefore we have $-\tau^* = -(\tau_2^* - \tau_1^*) > 0$. By constraint (1.9), to increase the consumption level of intervention j, θ_j, we should increase the utility $u_q(-e_j)u_v(-\beta_j p_j - c_j)$. By constraint (1.29), increasing $u_q(-e_j)u_v(-\beta_j p_j - c_j)$ by one unit increases the HP's objective function by τ_1^* and by constrain (1.30), increasing $u_q(-e_j)u_v(-\beta_j p_j - c_j)$ by one unit decreases the HP's objective function by τ_2^*. Therefore increasing $u_q(-e_j)u_v(-\beta_j p_j - c_j)$ by one unit results in increasing the HP's objective function by $-(\tau_2^* - \tau_1^*)$ which is positive. Thus $-\tau^* > 0$ implies a positive WTP for expansion. The case for $-\tau^* < 0$ is similar and hence omitted.

By constraint (1.9), to increase the consumption level of intervention j, θ_j, by one unit, we should increase the utility $u_q(-e_j)u_v(-\beta_j p_j - c_j)$ by $-(U_j'(\theta) - U_0'(\theta))$ unit. Consequently, increasing the consumption level of intervention j by one unit increase the HP's objective function by $-\tau^*\big(-(U_j'(\theta) - U_0'(\theta))\big)$, which is equal to $\gamma^* = \tau^* C_j'(\theta_j^*)$. □

A.3. Proof of Proposition 1

The results are immediate from the first-order conditions (1.13)-(1.16). We only show the first part; parts 2 and 3 are proved similarly. By Eq. 1.16, When $0 < \beta_j^* < \beta_j^{\max}$, we have $\delta^* = 0$; hence, by Eq. 1.13, we have $\theta_j^* + \tau^* u_q(-e_j)u_v'(-\beta_j^* p_j - c_j) = 0$. Consequently, by Theorem 2, the HP's WTP for expansion will be $\gamma^* = \tau^* C_j'(\theta_j^*)$, which results in Eq. 1.17.

To show Part 1.(b), By Eq. 1.15, when $0 < \theta_j^* < 1$, we have $\eta^* = 0$, which makes Eq. 1.14 equivalent to:

$$\pi_j(\theta_j^*, \beta_j^*) - \pi_0(\theta_j^*) - \tau \left.\frac{dC_j}{d\theta}\right|_{\theta=\theta_j^*} = 0. \tag{1.35}$$

Substituting $\pi_j(\theta, \beta_j) = (1-\beta)v_j(\theta) - (1-\beta_j)p_j + \lambda(q_j(\theta) - e_j)$ and $\pi_0(\theta) = (1-\beta)v_0(\theta) + \lambda q_0(\theta)$ in Eq. 1.35 results in Eq. 1.18. □

A.4. Proof of Proposition 2

Suppose that at optimality, there are M interventions in the optimum sequence s. The contract $(\beta_1, \ldots, \beta_M)$ divides the population into at most $M+1$ sections, each of which uses only one of the available interventions (See Figure 1.1(a)). Assume that the population is split at points $(\theta_1, \ldots, \theta_{M-1}, \theta_M)$, where $0 = \theta_{M+1} \leq \theta_M \leq \theta_{M-1} \leq \ldots \leq \theta_1 \leq \theta_0 = 1$. The HP solves the following problem to determine the coinsurance rates $(\beta_1, \ldots, \beta_M)$ and the consumption level $(\theta_1, \ldots, \theta_M)$ (recall that an individual of type $\theta \in (\theta_{j+1}, \theta_j)$ uses the intervention j):

$$\max_{\substack{\beta_1,\ldots,\beta_M \\ \theta_1,\ldots,\theta_M}} \Pi = \sum_{j=0}^{M} \int_{\theta_{j+1}}^{\theta_j} \pi_j(t, \beta_j)\, dt \tag{1.36}$$

$$\text{s.t. } u(-e_j, -\beta_j p_j - c_j) + U_j(\theta_j) = u(-e_{j-1}, -\beta_{j-1} p_{j-1}) + U_{j-1}(\theta_j), \text{ for } j = 1, \ldots, M, \tag{1.37}$$

$$0 = \theta_{M+1} \leq \theta_M \leq \theta_{M-1} \leq \ldots \leq \theta_1 \leq \theta_0 = 1, \tag{1.38}$$

$$0 \leq \beta_j \leq \beta_j^{\max}, \text{ for } j = 1, \ldots, M. \tag{1.39}$$

We solve the optimization problem (1.36)-(1.39) in M steps (stages). In stage k, given the optimal allocation of the intervention j_{k-1}, $\theta^*_{j_{k-1}}$ and ignoring the interventions $(j_{k+1}, \ldots, j_K)$, we determine the portion of population that should use the intervention j_{k-1} by solving:

$$\max_{\substack{0 \leq \beta_k \leq \beta_k^{\max} \\ 0 \leq \theta_k \leq \theta^*_{k-1}}} \int_0^{\theta_k} \left(\pi_k(t, \beta_k) - \pi_{k-1}(t, \beta^*_{k-1}) \right) dt \tag{1.40}$$

$$\text{s.t. } u(-e_j, -\beta_j p_j - c_j) + U_j(\theta) = u(-e_{j-1}, -\beta^*_{j-1} p_{j-1}) + U_{j-1}(\theta^*_j). \tag{1.41}$$

If the optimal allocation of the intervention k turns out to be $\theta^*_k = \theta^*_{k-1}$, then the intervention $k-1$ is dominated by interventions k and $k-2$, and hence should be dropped from the current sequence.

We claim that this approach is in fact solving the optimization problem (1.36)-(1.39) . To prove this, we show that (1) the objective function being maximized by this method is equal to the objective function (1.36), and (2) at each stage, the feasibility constraints (1.37)-(1.39) are not violated.

To show the first part, the first key observation is that by constraint (1.37), all β_k's can be eliminated from the optimization problem (1.36)-(1.39), and hence, at each stage, the HP's problem is to find only the *allocation* of each intervention, i.e., θ^*_k's. For instance, at stage $k = 1$, by solving (1.40)-(1.41) for $k = 1$, the HP determines the portion of population who uses the intervention $j = 0$ (use nothing), i.e. $[\theta^*_1, 1]$ (See Figure 1.1(b)); at stage $k = 2$, by solving

(1.40)-(1.41) for $k = 2$, the HP determines the portion of population who uses the intervention $j = 1$, i.e. $[\theta_2^*, \theta_1^*]$ (See Figure 1.1(c)), and so on.

The second key observation is that the optimal decision at stage k, i.e., θ_k^*, does not affect the value of the objective function (1.40) in problem (1.40)-(1.41) for the subsequent stages (stages $> k$). In other words, the decision at stage k does not impose a new bound on the objective function of the subsequent stages. To see this, recall that if the optimal allocation of intervention $k+1$ turns out to be $\theta_{k+1}^* = \theta_k^*$, then since the intervention k is dominated by interventions $k+1$ and $k-1$, we drop it from the sequence; and therefore, in the optimum sequence, we always have $\theta_{k+1}^* < \theta_k^*$, which means that in the optimum sequence, the decision in each stage, say θ_k^*, does not add a constraint to the decision in the subsequent stages, i.e., $\theta_j^*, j \geq k+1$.

And finally, we show that the summation of the objective functions (1.40) is equal to the objective function (1.36). Subtracting the constant $\int_0^1 \pi_0(t)dt$ from the objective function (1.36) does not change the optimal solution; therefore, maximizing Π is equivalent to maximizing:

$$\begin{aligned}
\Pi - \int_0^1 \pi_0(t)dt &= \sum_{j=0}^{M} \int_{\theta_{j+1}}^{\theta_j} \pi_j(t, \beta_j)\, dt - \int_0^1 \pi_0(t)dt \\
&= \sum_{j=0}^{M} \left(\int_0^{\theta_j} \pi_j(t, \beta_j)\, dt - \int_0^{\theta_{j+1}} \pi_j(t, \beta_j)\, dt \right) - \int_0^1 \pi_0(t)dt \\
&= \sum_{j=1}^{M} \int_0^{\theta_j} \big(\pi_j(t, \beta_j) - \pi_{j-1}(t, \beta_{j-1}) \big)\, dt. \qquad (1.42)
\end{aligned}$$

Eventually, according to the preceding argument, it is apparent that the optimal allocation obtained in each stage satisfies the feasibility constraints (1.37)-(1.39). □

A.5. Proof of Proposition 3

We write the equality constraint (1.20) as two inequality constraints:

$$U_{j_k}(\theta_{j_k}) - U_{j_{k-1}}(\theta_{j_k}) - u(-e_{j_{k-1}}, -\beta_{j_{k-1}}^* p_{j_{k-1}}) \geq -u(-e_{j_k}, -\beta_{j_k} p_{j_k} - c_{j_k}), \qquad (1.43)$$

$$U_{j_k}(\theta_{j_k}) - U_{j_{k-1}}(\theta_{j_k}) - u(-e_{j_{k-1}}, -\beta_{j_{k-1}}^* p_{j_{k-1}}) \leq -u(-e_{j_k}, -\beta_{j_k} p_{j_k} - c_{j_k}). \qquad (1.44)$$

We then proceed in the exact same way as in the proof of Theorem 2.

A.6. Continuous Approximation

In this section we present a methodology to determine the rank of each health profile and to approximate the the discrete functions $\mathrm{E}[\tilde{q}_j(\cdot)]$ and $\mathrm{E}[\tilde{v}_j(\cdot)]$ by continuous and differentiable

functions. We assume that the distribution of parameter α_i, which represents the proportion of the population with health profile ϕ_i, is known, and for each health profile ϕ_i, we can calculate $\mathrm{E}[\tilde{q}_j(\phi_i)]$ and $\mathrm{E}[\tilde{v}_j(\phi_i)]$. Let Φ denote the set of health profiles. To find the approximation functions $q_j(\cdot)$ and $v_j(\cdot)$, the main idea is to order the health profiles such that the functions $\mathrm{E}[\tilde{q}_j(\cdot)]$ and $\mathrm{E}[\tilde{v}_j(\cdot)]$ can be best fit by the continuous functions $q_j(\cdot)$ and $v_j(\cdot)$. To describe the methodology, let us begin with a special case where for each of two health profiles ϕ_i and $\phi_{i'}$, $\mathrm{E}[\tilde{v}_j(\phi_i)] \geq \mathrm{E}[\tilde{v}_j(\phi_{i'})] \Leftrightarrow \mathrm{E}[\tilde{q}_j(\phi_i)] \geq \mathrm{E}[\tilde{q}_j(\phi_{i'})]$. In this case, sorting based on either $\mathrm{E}[\tilde{q}_j(\phi_i)]$ or $\mathrm{E}[\tilde{v}_j(\phi_i)]$ results in the same ordering. Thus, we sort the health profiles in the decreasing order of either $\mathrm{E}[\tilde{q}_j(\phi_i)]$ or $\mathrm{E}[\tilde{v}_j(\phi_i)]$ to obtain the ordered set $\vec{\Phi}$. Over the ordered set $\vec{\Phi}$, we define the function $\Theta : \vec{\Phi} \to [0,1]$ as $\Theta(\phi_{(i)}) = \sum_{s=1}^{i-1} \alpha_{(s)} + \alpha_{(i)}/2$. Now we find the function $v_j(\cdot)$ that minimizes $SSE_v = \sum_{\phi_{(i)} \in \vec{\Phi}} \left(v_j(\Theta(\phi_{(i)})) - \mathrm{E}[\tilde{v}_j(\phi_{(i)})]\right)^2$ and the function $q_j(\cdot)$ that minimizes $SSE_q = \sum_{\phi_{(i)} \in \vec{\Phi}} \left(q_j(\Theta(\phi_{(i)})) - \mathrm{E}[\tilde{q}_j(\phi_{(i)})]\right)^2$.

In most practical cases, however, the assumption of this special case does not necessary hold. Therefore, to order the health profiles for general cases, we define a new measure $R : \Phi \to \mathbb{R}$ to represents the *importance* of each health profile from the insurer's perspective. We define the function $R(\cdot)$ as $R(\phi_i) = \mathrm{E}[\tilde{v}_j(\phi_i)] + \rho \mathrm{E}[\tilde{q}_j(\phi_i)]$, where ρ is a positive constant representing the relative importance of health to cost from the insurer's perspective. In §1.2.3, we showed that the insurer in fact evaluates each health profile by a linear combination of its prospective cost and QALY gain (or loss). Hence defining the importance function $R(\cdot)$ as a linear function is in line with our previous findings.

Next for a given value of ρ and the set of health profile Φ, we sort the health profiles in the decreasing order of their importance, i.e., $R(\phi_i)$, to obtain the ordered set $\vec{\Phi}$. Over the ordered set $\vec{\Phi}$, we defined the function $\Theta : \vec{\Phi} \to [0,1]$ as before: $\Theta(\phi_{(i)}) = \sum_{s=1}^{i-1} \alpha_{(s)} + \alpha_{(i)}/2$. The goodness-of-fits for the functions $q_j(\cdot)$ and $v_j(\cdot)$ depend on the ordering of the health profiles in set $\vec{\Phi}$, which depends on the value of ρ. To see the relation, for a given ρ, find the ordered set $\vec{\Phi}$; over the ordered set $\vec{\Phi}$, find the function $v_j(\cdot)$ that minimizes $SSE_v(\rho) = \sum_{\phi_i \in \vec{\Phi}} \left(v_j(\Theta(\phi_i)) - \mathrm{E}[\tilde{v}_j(\phi_i)]\right)^2$ and the function $q_j(\cdot)$ that minimizes $SSE_q(\rho) = \sum_{\phi_i \in \vec{\Phi}} \left(q_j(\Theta(\phi_i)) - \mathrm{E}[\tilde{q}_j(\phi_i)]\right)^2$. The optimal ρ is determined by: $\rho^* = \arg\min_{\rho}(SSE_v(\rho) + \rho^2 SSE_q(\rho))$. The functions $q_j(\cdot)$ and $v_j(\cdot)$ can now be estimated accordingly.

For the numerical application presented in §1.4 of the paper, we assumed that functions $q_j(\cdot)$ and $v_j(\cdot)$ have exponential form of $f(x) = a + be^{cx}$. Table 1.2 shows the estimated parameters for the cost and QALY resulted from each test. All the R^2's are greater than 90% implying that the fitted models in fact account for more than 90% of the variation in the observed data.

Table 1.2: Parameter estimation for treatment cost and QALY gain

Alternatives	NO		FOBT		SIG		COL	
	Cost	QALY	Cost	QALY	Cost	QALY	Cost	QALY
a	77.695	0.066	87.616	0.064	72.562	0.067	63.218	0.080
b	388.540	0.078	330.030	0.075	312.495	0.069	248.777	0.056
c	7.622	1.373	7.575	1.431	8.124	1.502	6.581	2.499
R^2	0.954	0.950	0.962	0.921	0.956	0.907	0.973	0.904

NO: No screening FOBT: Fecal Occult Blood test SIG: Sigmoidoscopy COL: Colonoscopy
All costs and QALYs are discounted at 3% rate to year 2005.

A.7. Screening Regimens

The latest update on recommended screening tests and the screening intervals are provided in the 2008 joint guideline from American Cancer Society, US Multi-Society Task Force on Colorectal Cancer, and the American College of Radiology (Levin et al. 2008). The guideline recommends for an average risk population to obtain one of the following CRC screening regimens: annual FOBT, Sigmoidoscopy every 5 years, or Colonoscopy every 10 years. For population at increased risk, undergoing Colonoscopy every 5 years is recommended regimen. The guideline defines the population at increased risk as having either colorectal cancer or adenomatous polyps in a first-degree relative before age 60 years or in 2 or more first-degree relatives at any age.

For a person who chooses FOBT, if the FOBT result is positive, the guideline recommends obtaining a Colonoscopy as a more accurate test, in the same year. If a patient with positive FOBT result chooses not to use the recommended Colonoscopy, she will retake FOBT in one year. If the FOBT result is negative, the follow-up FOBT for a person with no history of adenoma or CRC is scheduled in 2 years. If the Colonoscopy result is positive, a Polypectomy (Biopsy) will be performed and the person goes under surveillance. Surveillance refers to the regular examination of the colon in patients with a prior history of colorectal adenoma or cancer history. Because a history of colorectal adenomas or cancer is a risk factor for future colonic neoplasia, surveillance regimens are more aggressive than standard screening strategies. Therefore, the frequency of the screening tests should be increased for a person under surveillance. If Colonoscopy finds advanced adenomas, or 3 or more non-advanced adenomas, the next Colonoscopy will be scheduled in 3 years after the initial Colonoscopy; and if the result is normal, the procedure will be repeated in 5 years. If the Colonoscopy finds non-advanced adenomas, the next Colonoscopy will be scheduled in 5 years after the initial Colonoscopy; and if the result is normal, the procedure will be repeated in 5 years. For a patient with a history of resection or colorectal cancer, the next Colonoscopy will be scheduled within 1 year of cancer resection; if the exam is normal, the procedure will be repeated in 3 years and if still normal, it

will be repeated in 5 years. Surveillance tests halt after the cancer goes into the terminal stage or the patient reaches age 80.

If a person decides not to use any screening tests and becomes aware of the cancer through its symptoms, a Colonoscopy is performed to determine the extent of the cancer, and treatment of the cancer begins. As part of Colonoscopy, adenomas may be found and removed. Should surgery prove necessary, sections of the colon are resected and thus eliminating all the adenomas in those sections.

A.8. Sensitivity Analysis

Our model estimates the WTP for health in a framework modeled as a game between the health insurer and the population. Hence, the estimated insurer's WTP for health is expected to be sensitive to the population's parameters, i.e., the population's health and monetary risk aversion (r_q and r_v, respectively). As discussed in §1.4.3, we chose r_q and r_v such that the sum of squared errors in problem (1.24) becomes minimum. For our defined population, the resulting r_q and r_v yields the average WTP of \$28 for FOBT and \$48 for Sigmoidoscopy. These results are in line with the findings of Frew et al. (2001) who estimate the WTP for CRC screening tests to be $\$30 - \$50$, independent of the screening protocol. Figure 1.6 depicts how the estimated health insurer's WTP for health changes with r_q and r_v. This figure shows that the estimated WTP for health is rather sensitive to the population's monetary and health risk aversion. Moreover, Figure 1.7 reveals that the sum of squared errors of problem (1.24) is not extremely sensitive to r_q and r_v, and hence this measure cannot perfectly detect the departures from the true values of r_q and r_v. Therefore, in our example when both r_q and r_v are unknown, extreme care should be taken when estimating the WTP for health, since inaccurate estimates for r_q and r_v may lead to imprecise estimate for the WTP for health. Nonetheless, for many diseases, there often exist empirical studies on the population's WTP for the related interventions, from which the experimenter can obtain an approximate estimates for the population's health and monetary risk aversion, as we did for the CRC screening tests.

Figure 1.8 shows how the estimated WTP for health changes with the observed consumption level.

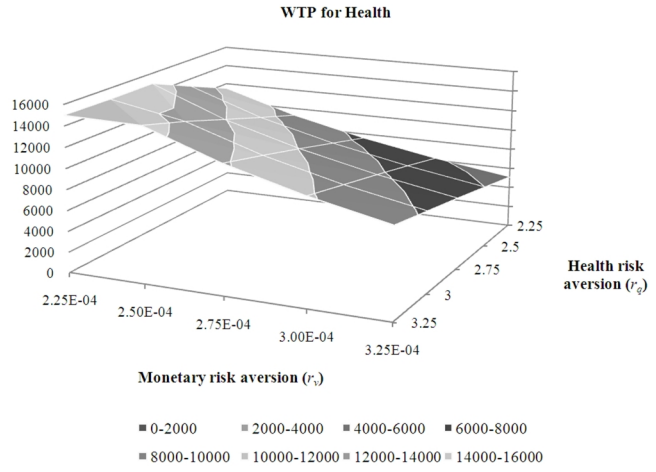


Figure 1.6: WTP for health versus the health and monetary risk aversion

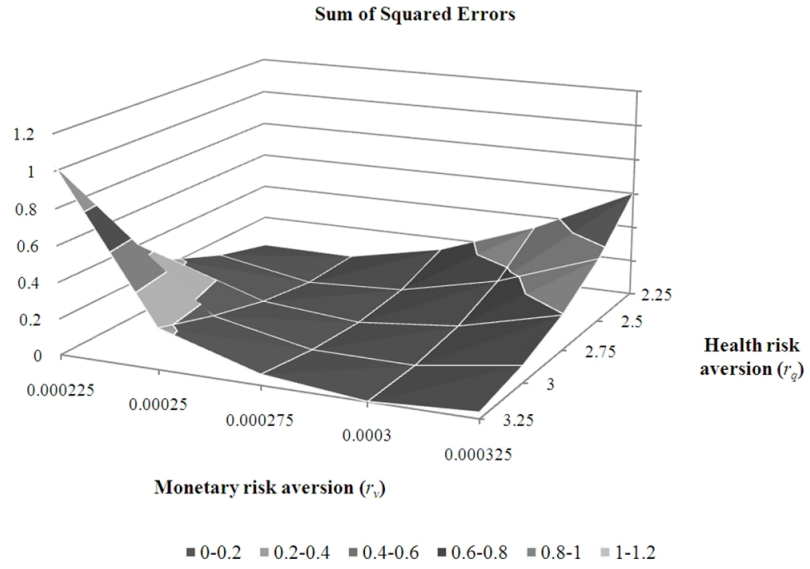


Figure 1.7: Sum of Squared Errors versus the health and monetary risk aversion

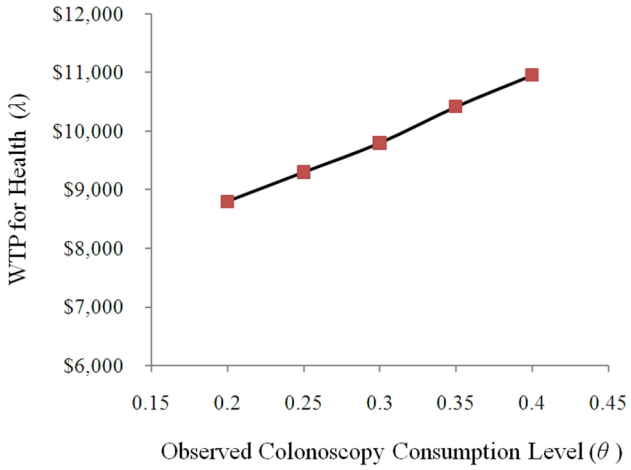


Figure 1.8: WTP for health versus the observed consumption level

Chapter 2

Coordinating Contracts in a Preventive Health Care System

Abstract

We consider a health care system consisting of two noncooperative parties: a health purchaser (e.g., a health insurer) and a health provider (e.g., a hospital). A principal-agent model is proposed to capture the interaction between the two parties. In this model, the health provider determines the type of patients who need a preventive medical intervention, and gets reimbursed by the health purchaser based on the number of patients for whom the intervention is administered. We determine the contracts that *coordinate* the health purchaser-health provider relationship; i.e., the contracts that allow each entity to optimize its own objective function while maximizing the population's welfare. We characterize the coordinating contracts for two settings: we show that under certain conditions (1) when the number of customers for the medical intervention is verifiable, there exist a gate-keeping contract and a set of concave piecewise linear contracts that coordinate the system; and (2) when the number of customers is not verifiable, there exist a contract of bounded linear form and a set of incentive-feasible concave piecewise linear contracts that coordinate the system. We also characterize the incentive feasible menu of coordinating contracts for cases where the health providers are located in communities at different risk of developing the disease.

Key words: coordinating contracts, health care, payment systems, mechanism design, principal-agent models.

2.1 Introduction

Suboptimal assignment of patients to medical interventions is one of the major sources of inefficiency in health care systems. Such assignments, which are usually determined by physicians and hospitals, can be rectified in part by designing *coordinating contracts* between the health purchaser and the health providers, in order to improve the efficiency of the health care system and the population's welfare.

The purchaser-provider relationship is prevalent in health care delivery systems. In this relationship, a health purchaser (e.g., health insurer) and a health provider (e.g., hospital) enter into a contractual agreement in which the health provider agrees to deliver service to the population it covers, and be reimbursed by the health purchaser according to a prespecified contract (payment system). A contract is said to *coordinate* a health care system if the set of health care system optimal actions is a Nash equilibrium, i.e., no entity in the system has a profitable unilateral deviation from the equilibrium that is optimal for the health care system.

In this paper, we consider a health care system consisting of a health purchaser, health providers, and a population. The population is assumed to be at risk of contracting or developing a disease for which one preventive medical intervention is available. The population is assumed to be *heterogeneous*, whose individuals are at different risk of developing the disease or prospective devastation caused by the disease. To each individual, we assign a number, called "rank," based on her expected level of devastation due to the disease, if once developed. During a contractual period, a random number of individuals visit the health provider to obtain the preventive medical intervention, if prescribed by the health provider. Having observed the health purchaser's contract, the health provider specifies a *threshold* (hidden from the health purchaser), and administers the intervention to individuals of rank lower than the threshold. The health purchaser then reimburses the health provider based on the number of people who used the preventive intervention during the contractual period.

To characterize the coordinating contracts, we first assume that during the contractual period, the number of customers for the underlying medical intervention is *verifiable* by the health purchaser; that is, the health purchaser can obtain the exact number of individuals who visited the health provider to obtain the medical intervention. This assumption may be valid in unsophisticated health care environments (for instance, when the health provider offers only one medical intervention) or in the presence of an advanced information system. Nonetheless, as will be discussed in §2.4, for rather complex health care delivery environments and in the absence of a proper information system, this assumption may be undermined. For these situations, however, we assume that at the beginning of each contractual period, both the health purchaser and the health provider can obtain a prior distribution for the number of customers for the underlying medical intervention during the contractual period.

We model the interaction between the health purchaser and the health provider as a principal-agent model. Our model is a mixed model of moral hazard and asymmetric information: it is a problem of moral hazard since the threshold specified by the health provider is hidden from the health purchaser; and it is a problem of asymmetric information since the patient texture (i.e., the distribution of the risk categories[1] in a heterogeneous population) served by each health provider is not observable by the health purchaser. Also, when the number of customers is not verifiable, the health purchaser is faced with two-dimensional asymmetric information caused by unobservable number of customers and the population's texture. Assuming that the health purchaser maximizes the population welfare while the health provider maximizes his profit, we derive coordinating contracts that aligns the health purchaser-health provider relationship.

2.1.1 Motivating Example

This work is primarily motivated by the important role of insurer-provider contracts in the utilization of cancer screening tests, and in particular Colorectal Cancer screening tests. Like many other cancers, Colorectal Cancer (CRC) becomes symptomatic when the cancer is at an advanced level and consequently the chance of survival is significantly low. However, the CRC screening tests can identify the cancer at an earlier stage leading to improved survival and considerably lower treatment cost.

There are several screening tests available for CRC, including Fecal Occult Blood Test (FOBT), Flexible Sigmoidoscopy, Combination of FOBT and Flexible Sigmoidoscopy, and Colonoscopy. The CRC screening tests have different performance in terms of cost and effectiveness, which are well examined in the literature (for instance, refer to Levin et al. (2008)). The preferred screening test for each individual is determined according to a number of risk factors, such as age, race, gender, and family history of CRC. For instance, for those with family history of CRC or previous incidence of adenoma, Colonoscopy is the preferred screening alternative, since it yields the best outcomes in terms of costs and life-years saved for this risk category (Levin et al. 2008). Nonetheless, the patient preference for CRC screening has been shown to be strongly sensitive to out-of-pocket costs, including copayments and the indirect costs of obtaining the screening tests (Pignone et al. 1999), and therefore, many individuals may decide not to use any screening tests or may select the one that yields an inferior outcome (Swan et al. 2003, American Cancer Society 2007).

An immediate solution for improving the cancer prevention and early detection practice

[1] In the context of preventive care, an individual's risk of developing a disease is usually determined by a number of "risk factors." For instance, the risk factors affecting the probability of developing a Colorectal cancer are age, race, gender, previous history of adenomas, and family history of cancer. Each combination of risk factors constitute a "risk category." For example, a risk category for Colorectal Cancer can be white males at age 60 without family history of cancer and adenomas.

is to reduce or eliminate the co-payments for preventive screening tests (Harvard Center for Cancer Prevention 2008, Ward et al. 2008). This strategy, however, is not necessarily sustained by health purchasers who generally believe that reducing the coinsurance rates aggravates the moral hazard effect in health care markets. In our screening example, the concern about moral hazard is rooted in the way that the health purchaser reimburses the health provider for the service rendered. The reimbursement system for screening tests is typically *fee-for-service*, in which the health provider receives a flat rate payment from the health purchaser for the service provided to the patient. Under this payment system, the health provider has the maximum incentive to screen *anyone* who is willing to pay for the copayment, leading to overconsumption of the screening tests. Hence, the health purchaser tends to maintains the coinsurance rates as a means to avoid such *inefficiency*.

Besides leading to an inefficient allocation, this mechanism also results in an *inequitable* allocation. That is, many people are not screened, not because they do not benefit from the tests but because they could not afford the copayments, and likewise, many people are screened simply because they are *induced* to undergo them, despite not benefiting much from the screening tests.

2.1.2 Generality

The inefficiency and inequity raised by the mechanism described above is prevalent in health care systems. The contracts offered to health providers may induce overutilization (or underutilization in some cases), and hence the health purchaser attempts to control the consumption by copayments. In consequence, many individuals do not receive the care they need because the copayments are not affordable for them.

Moreover, in a health care market, health providers may have characteristics that are not visible to the health purchaser; for instance, different health providers have different cost structures and may serve different patient textures. These types of heterogeneity in health providers makes designing the contracts even more complex, since under these conditions the data needed to enforce a particular contract becomes prohibitively costly to collect.

The above discussion motivates us to pose a research question: What contracts allow both the health purchaser and the health provider to optimize their own objective functions while maximizing the population's welfare? In the context of operations management, such contracts are generally referred to as "coordinating contracts" (Cachon 2003); those are the contracts that results in Nash equilibrium while optimizing the global system.

2.1.3 Research Themes

Several reimbursement systems have been proposed and employed for health care systems. In the e-companion attached to this paper, we have discussed each payment system briefly, and have reviewed some recent work in this area. The reader is referred to Newhouse (1996) for a more thorough overview of the health care reimbursement systems. In these studies, the interactions between the health purchaser and health provider are modeled in principal-agent frameworks, where a principal (e.g., health purchaser) delegates a task (e.g., providing medical service to the population) to an agent (e.g., health provider). The principal's problem is to design a contract that induces the agent to take the action(s) desired by the principal.

One theme of research focuses on the inefficiencies caused by *hidden information*, e.g., the purchaser has imperfect knowledge about the cost structure of the provider. These studies aim at finding the incentive-compatible contracts to reduce such inefficiencies (Boadway et al. 2004, Jack 2005, Shleifer 1985). Another stream of research focuses on the *hidden action* problem, which occurs when the purchaser cannot observe the provider's action, such as, the *intensity* of treatment (Fuloria and Zenios 2001, Ma and McGuire 1997) or the *quality* of the delivered care (Jack 2005, Chalkley and Malcomson 1998, Ma 1994). Monitoring the quality of treatment and the physician's effort are either not possible or too costly to be worthwhile, and hence usually treated as hidden action.

When the main objective is to optimize the *global* system, the problem of moral hazard and asymmetric information can be accommodated through "coordinating contracts." Coordinating contracts have gained significant attention in the operations management literature (for a comprehensive review, refer to Cachon (2003)), and different types of coordinating contracts have been introduced and studied; for instance, see wholesale price contracts (Cachon 2003), payback contracts (Pasternack 1985), revenue-sharing contracts (Cachon and Lariviere 2005), and quantity-discount contract (Weng 1995). Nonetheless, most of these contracts are developed specifically for manufacturer - supplier or supplier - retailer paradigms, with very limited ramifications for health care payment systems.

The health provider in our setting may also be considered as a "gate keeper" who refers the patients at higher risk to a specialist to receive the preventive intervention. Gate keeper's contracts usually assume that the number of customers for the medical intervention is verifiable; and therefore these contracts reimburse the health provider based on the number of referrals and the number of patients treated without referral; for instance see Shumsky and Pinker (2003) and Marioso and Jelovac (2003). In §2.3, where we assume that the demand verifiability assumption holds, we show that gate keeper's contracts can coordinate the system under certain conditions.

This paper makes the following contributions: (1) it develops a set of novel health care reimbursement systems that leads to maximizing the social welfare, while at the same time

allowing each market entity to optimize their own objective functions; and (2) contrary to the common health care reimbursement systems, which only consider the problem of hidden action or hidden information, our proposed principal-agent model is a mixed paradigm of moral hazard (hidden effort level) and two-dimensional asymmetric information (hidden number of customers and the population's texture).

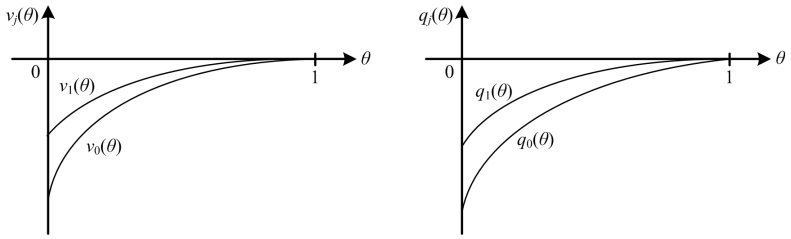
The remainder of this paper is organized as follows: Section 2.2 details the model and defines the coordinating contracts in the underlying system. In §2.3, we characterize the coordinating contracts and formulate a principal-agent model to determine a set of incentive-compatible coordinating contracts to be offered to the health providers with different patient textures. The models proposed in §2.3 assume that the number of customers is verifiable; in §2.4, we relax this assumption and find the coordinating contracts when the number of customers for the medical intervention is not verifiable by the health purchaser. Section 2.5 discusses future extensions and concludes the paper.

2.2 The Model and the Coordinating Contracts

We consider a population which is at risk of contracting a disease. The population can undergo a preventive intervention, at a price, in order to reduce the risk of developing the disease or to alleviate the devastation caused by the disease once developed. The population is assumed to be *heterogeneous* in which individuals have different expected magnitude of suffering from the disease. Hence, not all individuals benefit from the preventive intervention to the same level; and the health purchaser, depending on her willingness-to-pay for health, may not be willing to pay for those at trivial risk.

We assume that the individuals can be ranked in a decreasing order of the expected devastation due to the disease, and a continuous value $\theta \in [0, 1]$ can be assigned to each individual such that $100\theta\%$ of population expect higher magnitude of suffering from the disease. The suffering from a disease may be the result of a financial loss or health deterioration. We assume that the financial loss is equal to the necessary treatment costs; in the presence of universal health care, however, other financial burden imposed to the society may be also included as "financial loss." We further assume that health can be quantified by a measure such as Quality-Adjusted Life Year (QALY) which is an aggregate variable to represent both the length of life and the health quality of years lived (Miyamoto et al. 1998, Pliskin et al. 1980).

Let $q_j(\cdot)$ be a function returning the expected health gain (or loss) for a patient of rank θ who uses alternative $j \in \{0, 1\}$, where $j = 0$ and $j = 1$ denote 'not using' and 'using' the intervention, respectively. The function $q_j(\cdot)$ may include losses that might occur due to the disease, normal aging, or an invasive intervention. Let $v_j(\cdot)$ be a function returning the expected treatment costs for a patient of rank θ who uses alternative $j \in \{0, 1\}$, excluding the price of



(a) Cost functions of Assum. 2 (b) Health functions of Assum. 3

Figure 2.1: Cost and health functions

the medical intervention obtained.

To offer the intervention, the health provider incurs a variable cost c and a fixed cost F. The fixed cost F may also include the reservation utility of the health provider. If the health provider decides to administer the intervention to the individuals of rank less than or equal to θ, then he incurs a cost (or disutility) $d(\theta)$, with $d'(\theta) > 0$ and $d''(\theta) > 0$. The cost (or disutility) $d(\cdot)$ occurs mainly due to (1) administering the intervention to a larger portion of the population requires more medical and financial resources to be allocated, which increases the cost of lost opportunities for the health provider; and (2) as an individual's risk of contracting the disease reduces, it takes more effort to persuade her to use the intervention.[2]

Let λ denote the health purchaser's willingness-to-pay (WTP) for health, and define the function $\pi_j(\cdot)$ as $\pi_j(\theta) = v_j(\theta) + \lambda q_j(\theta)$ for $j \in \{0, 1\}$. The function $\pi_j(\cdot)$ returns the health purchaser's net monetary benefit if the patient of rank θ uses the alternative $j \in \{0, 1\}$. Define $\Pi(\theta) = \int_0^\theta (\pi_1(t) - \pi_0(t))\, dt$ as the health purchaser's gain if individuals of rank θ and less undergo the intervention. We assume that the functions $v_j(\cdot)$ and $q_j(\cdot), j \in \{0, 1\}$, satisfy the following unrestrictive assumptions such that the function $\pi_1(\theta) - \pi_0(\theta)$ is decreasing in θ; and therefore, $\Pi'(\cdot) > 0$ and $\Pi''(\cdot) < 0$.

Assumption 2 *For any $\theta \in [0, 1]$, (1) $v_j'(\theta) \geq 0$, for $j = 0, 1$, i.e., for each alternative, as θ increases (lowering the expected magnitude of suffering), the expected treatment costs reduces; and (2) $v_1'(\theta) \leq v_0'(\theta)$, that is, the effectiveness of the intervention in reducing the treatment cost (i.e., $v_1(\theta) - v_0(\theta)$) decreases in θ.*

Figure 2.1(a) shows two functions that satisfy the requirements in Assumption 2.

Assumption 3 *For any $\theta \in [0, 1]$, (1) $q_j'(\theta) \geq 0$, for $j = 0, 1$, i.e., for each alternative, as θ increases (lowering the expected magnitude of suffering), the expected loss in health reduces; and (2) $q_1'(\theta) \leq q_0'(\theta)$, i.e., the effectiveness of the intervention in reducing the expected health loss (i.e., $q_1(\theta) - q_0(\theta)$) decreases in θ.*

[2] A methodology for estimating the function $d(\cdot)$ is proposed in the attached e-companion.

Figure 2.1(b) shows two functions that satisfy the requirements of Assumption 3.

Let the random variable N denote the number of customers for the medical intervention who visit the health provider during the contractual period. To be more precise, we define "the number of customers for a medical intervention, N," as the number of individuals who visit the health provider to consume the medical intervention if prescribed by the health provider. For instance, during an epidemic period, N would be the number of individuals who consider themselves at risk of contracting the disease and visit the health provider in order to seek the preventive procedure. Among these N individuals, the health provider then administers the intervention to those at higher risk of developing the disease. Let N have probability density $f_N(\cdot)$ and support $[\underline{n}, \bar{n}]$.

If the health provider administers the intervention to individuals with rank less than θ, we assume that the social welfare can be represented by:

$$\text{Social Welfare} = (\Pi(\theta) - c\theta - d(\theta))\,\mathrm{E}[N] - F. \tag{2.1}$$

Note that in social welfare (2.1), the disutility $d(\theta)$ is assumed to deteriorate the social welfare proportional to $\mathrm{E}[N]$. This assumption can be justified by the following argument. As described before, the disutility function $d(\cdot)$ consists of (1) the cost of lost opportunities incurred by the health provider, and (2) the cost of effort to persuade the population to use the intervention. For both cases, it is reasonable to assume that these components of disutility $d(\cdot)$ increase linearly by the population size.

To maximize the social welfare, the health purchaser maximizes (2.1) over θ. Since the social welfare (2.1) is concave in θ, the global optimal solution, $\tilde{\theta}$, only needs to satisfy the following first order condition, given that $0 < \tilde{\theta} < 1$:

$$\Pi'(\tilde{\theta}) = c + d'(\tilde{\theta}). \tag{2.2}$$

Therefore, to maximize the social welfare, the individuals of rank $\theta \in [0, \tilde{\theta}]$ should use the intervention and the individuals of rank $\theta \in (\tilde{\theta}, 1]$ should not use the intervention. Throughout this paper, we assume that the social welfare (2.1) is always greater than zero at the welfare maximizing threshold $\tilde{\theta}$. Greater $\tilde{\theta}$ implies that a larger percentage of population should undergo the intervention, which requires greater implementation *effort*. Therefore, to conform to contract theory literatures, we refer to θ as the *effort* to be exerted by the health provider in implementing the intervention.

In maximizing the social welfare (2.1), we assumed that the health purchaser has perfect information about the rank of each individual, θ, and that the effort level $\tilde{\theta}$ exerted by the health provider is observable. In practice, however, it is only the health provider who can observe the true rank of each patient and make the recommendations accordingly. It would be

too expensive for the health purchaser to obtain the true value of all the patients' rank (θ's) for whom the intervention has been prescribed. Therefore, the health purchaser's problem is to design a contract that leads the health provider to choose the welfare maximizing effort level $\tilde{\theta}$ as his optimal strategy. The contracts that allow both the health purchaser and the health provider to optimize their objective functions while implementing the social welfare maximizing effort level $\tilde{\theta}$ are called the *coordinating contracts*.

In finding the coordinating contracts, we also make the following assumption about the health provider throughout this paper.

Assumption 4 *The health provider is altruistic; that is, if for a given contract, exerting the welfare maximizing effort level $\tilde{\theta}$ is optimal for the health provider, then he administers the intervention to an individual of rank θ if and only if $\theta \leq \tilde{\theta}$.*

2.3 Coordinating Contracts with Verifiable Number of Customers for the Medical Intervention

In this section, we assume that the number of customers for the underlying medical intervention who visit the health provider, N, is verifiable by the health purchaser. For instance, consider a temporary health care facility offering a single preventive intervention (such as a vaccine) during an epidemic period. The individuals who consider themselves at risk of contracting the disease visit the health provider, and the health provider administers the vaccine only to those with rank lower than the prespecified threshold. In such systems, the health purchaser can easily verify the number of customers for the preventive intervention by counting the total number of visits to the health provider. This assumption may also be valid when the health purchaser benefits from an advanced information system through which she can distinguish among the customers for different medical services.

2.3.1 Coordinating Contract with Homogenous Health Providers

In this subsection, we assume that the health purchaser knows the health provider's patient texture, i.e., the distribution of the risk categories in a heterogeneous population is known to the health purchaser. Let $w_n(\cdot)$ denote the health purchaser's reimbursement contract when the realized number of customers is n. Throughout this section, the subscript n in $w_n(\cdot)$ is to emphasize the verifiability of the number of customers, and the fact that the reimbursement $w_n(z)$ depends on both the realized number of customers, n, and the number of patients who used the intervention, z. When the number of customers is verifiable, we denote the health purchaser's contract by $\{w_n(\cdot)\}$, since it is in fact a collection of several functions $w_n(\cdot)$, each of which is designed for a particular realized number of customers, n.

Let θ_n denote the health provider's effort level when demand n is observed. Then, having observed the demand n, the health provider solves the following optimization problem to determine the effort level θ_n:

$$\max_{0\leq\theta_n\leq 1} w_n(\theta_n n) - c\theta_n n - nd(\theta_n) - F. \tag{2.3}$$

The health purchaser then solves the following program to find the coordinating contracts $\{w_n(\cdot)\}$ and the optimal allocations θ_n:

$$\max_{\{w_n(\cdot)\},\theta_n} \quad \int_{\underline{n}}^{\bar{n}} \big(n\Pi(\theta_n) - w_n(\theta_n n)\big) f_N(n)dn \tag{2.4}$$

$$\text{s.t.} \quad \theta_n = \arg\max_{0\leq\theta\leq 1} w_n(\theta n) - c\theta n - nd(\theta), \tag{2.5}$$

$$\int_{\underline{n}}^{\bar{n}} \big(w_n(\theta_n n) - c\theta_n n - nd(\theta_n)\big) f_N(n)dn \geq F. \tag{2.6}$$

The constraint (2.5) is the optimality condition of the problem (2.3), and the *participation* constraint (2.6) ensures that the contract $\{w_n(\cdot)\}$ compensates the health provider's expenses.

In this section, since we assume that the number of customers for the medical intervention is verifiable, we can employ a gate keeping contract (Shumsky and Pinker 2003, Marioso and Jelovac 2003) to coordinate the system. Consider a contract by which the health provider receives a payment w for each individual to whom the intervention is administered, and payment b for each individual who does not use the intervention. Therefore, $w_n(z) = wz + b(n-z), 0 \leq z \leq n$. The following proposition shows the conditions under which this contract coordinates the system.

Proposition 4 *When the number of customers is verifiable and the disutility function $d(\cdot)$ is strictly convex, the contract by which the health provider receives a payment w for each individuals to whom the intervention is administered, and a payment b for each individual who does not use the intervention, coordinates the system if:*

$$w = \frac{F}{\mathrm{E}[N]} + c + d'(\tilde{\theta}) + d(\tilde{\theta}) - \tilde{\theta}d'(\tilde{\theta}), \text{ and}$$
$$b = \frac{F}{\mathrm{E}[N]} + d(\tilde{\theta}) - \tilde{\theta}d'(\tilde{\theta}).$$

Proof. See Appendix □

Note that since $d(\cdot)$ is convex, the term $d(\tilde{\theta}) - \tilde{\theta}d'(\tilde{\theta})$ is always negative, and therefore, the payment b determined by Proposition 4 may become negative. A contract with negative payment may not be necessarily attractive to implement. The following proposition presents an alternative form of contract to avoid such problem.

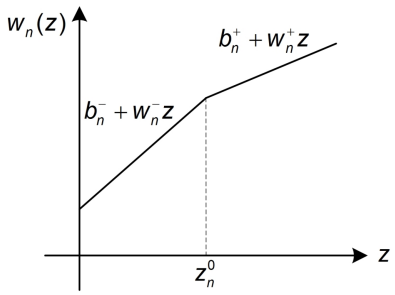


Figure 2.2: A concave piecewise linear contract

As an alternative, we consider a set of contract $\{w_n(\cdot)\}$ in which $w_n(\cdot)$'s are assumed to be concave piece-wise linear contracts of form:

$$w_n(z) = \begin{cases} b_n^- + w_n^- z, \text{ for } z \leq z_n^0, \\ b_n^+ + w_n^+ z, \text{ for } z \geq z_n^0, \end{cases} \tag{2.7}$$

where $w_n^- < w_n^+$.

Figure 2.2 displays a piecewise linear contract of form (2.7).

Proposition 5 *When the number of customers is verifiable, there exists a set of concave piecewise linear contracts $\{w_n(\cdot)\}$ of form (2.7) that coordinate the health purchaser-health provider relationship. For any realization of N, say n, the contract $w_n(\cdot)$ satisfies the following properties:*

$$\begin{aligned} z_n^0 &= \tilde{\theta} n, \\ w_n^- &= c + d'(\tilde{\theta}) + \epsilon, \\ w_n^+ &= c + d'(\tilde{\theta}) - \epsilon, \\ b_n^- &= F + n(d(\tilde{\theta}) - \tilde{\theta} d'(\tilde{\theta})) - \tilde{\theta} n \epsilon, \\ b_n^+ &= b_n^- + (w_n^- - w_n^+)\tilde{\theta} n, \end{aligned}$$

where $\tilde{\theta}$ satisfies Eq. 2.2 and $\epsilon > 0$ is large enought to make $w_n(\cdot)$ concave for $n \in [\underline{n}, \bar{n}]$.

Proof. The proof is similiar to that of Proposition 4 and hence omitted. □

2.3.2 Coordinating Contract with Heterogeneous Health Providers

In §2.3.1, we assumed that the health purchaser knows the health provider's patient texture. In reality, however, the health providers are generally located in communities with different

population risk distributions. In this section, we consider two health providers, one in a high-risk population and the other in a low-risk population. The health purchaser is not able to identify the type of each health provider since she does not possess accurate information about the patients served by each health provider. Nonetheless, she knows that a health provider is low-risk with probability α_L and high risk with probability α_L, where $\alpha_L + \alpha_H = 1$.

To prescribe the intervention for the patients of rank θ and lower, the health provider in low (high) risk population incurs disutility $d_L(\theta)$ ($d_H(\theta)$), where $d_H(\theta) < d_L(\theta)$, and $d'_H(\theta) < d'_L(\theta)$ for any $\theta \in [0,1]$. Define $\Pi_L(\theta_L) = \int_0^{\theta_L} \big(\pi_1^L(t) - \pi_0^L(t)\big)dt$ and $\Pi_H(\theta_H) = \int_0^{\theta_H} \big(\pi_1^H(t) - \pi_0^H(t)\big)dt$ as the health purchaser's gain if the individuals of rank lower or equal to θ_L in low risk population and of rank lower or equal to θ_H in high risk population undergo the intervention. It is reasonable to assume that $\Pi_L(\theta) < \Pi_H(\theta)$ and $\Pi'_L(\theta) < \Pi'_H(\theta)$ for any $\theta \in [0,1]$.

If the health purchaser knows the type of each health provider, then she can use Proposition 4 or Proposition 5 to specify the first-best contracts $\{w_{L,n}^{FB}(\cdot)\}$ and $\{w_{H,n}^{FB}(\cdot)\}$ to be offered to the low and high-risk health providers. Let $\tilde{\theta}_L^{FB}$ and $\tilde{\theta}_H^{FB}$ denote the socially optimal effort levels under complete information (first-best).

Lemma 2 *Under complete information,* $\tilde{\theta}_L^{FB} < \tilde{\theta}_H^{FB}$.

Proof. Suppose the contrary; i.e. $\tilde{\theta}_L^{FB} \geq \tilde{\theta}_H^{FB}$. Since $d_L(\cdot)$ is convex we have: $d'_L(\tilde{\theta}_L^{FB}) \geq d'_L(\tilde{\theta}_H^{FB})$. And Since $d'_L(\cdot) \geq d'_H(\cdot)$ by assumption, we will have $d'_L(\tilde{\theta}_L^{FB}) \geq d'_L(\tilde{\theta}_H^{FB}) \geq d'_H(\tilde{\theta}_H^{FB})$. Therefore, by Eq. 2.2 and the assumption that $\Pi'_L(\cdot) < \Pi'_H(\cdot)$, we should have $\Pi'_L(\tilde{\theta}_L^{FB}) \geq \Pi'_H(\tilde{\theta}_H^{FB}) > \Pi'_L(\tilde{\theta}_H^{FB})$. The inequality results in $\tilde{\theta}_L^{FB} < \tilde{\theta}_H^{FB}$, since $\Pi_L(\cdot)$ is concave. This contradicts the original assumption. □

Lemma 2 implies that for the high-risk population, the society benefits more if a larger portion of the population undergo the intervention, compared with the low risk population. Nonetheless, if the health purchaser offers the first-best contracts $\{w_{L,n}^{FB}(\cdot)\}$ and $\{w_{H,n}^{FB}(\cdot)\}$ to the low and high-risk health providers, the high-risk health provider may have an incentive to mimic the low-risk health provider by selecting the contract $\{w_{L,n}^{FB}(\cdot)\}$, since then, he incurs less disutility (recall that $d_H(\theta) < d_L(\theta)$ for any $\theta \in [0,1]$). In consequence, the high-risk health provider chooses a suboptimal effort level $\tilde{\theta}_L^{FB}$ instead of $\tilde{\theta}_H^{FB}$. To find the coordinating contracts under asymmetric information, we proceed in two steps:

Step 1: We first apply the Extended Revelation Principle (Laffont and Martimort 2001): according to this theorem, the principal can confine himself to offer a *truthful direct revelation mechanism* $\{w_{L,n}(\cdot), w_{H,n}(\cdot)\}$ and to recommend a choice of efforts θ_L^* and θ_H^*. Under this mechanism, the health purchaser only need to offer two sets of contracts $\{w_{L,n}(\cdot)\}$ and $\{w_{H,n}(\cdot)\}$ for low-risk and high-risk health providers, respectively, while assuring that health providers of each type reveals his type truthfully by selecting the contract

specifically designed for him. The result of this step is the optimal effort levels θ_L^* and θ_H^* by low-risk and high-risk health providers, respectively, and the expected amount of associated reimbursements, w_L^* and w_H^*.

Step 2: Given the optimal efforts θ_L^* and θ_H^*, and the expected reimbursement amounts w_L^* and w_H^*, we then specify the coordinating contracts $\{w_{L,n}(\cdot)\}$ and $\{w_{H,n}(\cdot)\}$ such that the health providers find exerting the effort levels θ_L^* and θ_H^* optimal.

Let $W_k(\theta) = \int_{\underline{n}}^{\bar{n}} w_{k,n}(\theta n) f_N(n) dn$, for $k \in \{L, H\}$. In Step 1, the health purchaser solves the following problem.

$$\max_{\substack{0\le\theta_L\le 1, W_L(\cdot)\\ 0\le\theta_H\le 1, W_H(\cdot)}} \quad \Pi = \alpha_L\big(\Pi_L(\theta_L)\mathrm{E}[N] - W_L(\theta_L)\big) + (1-\alpha_L)\big(\Pi_H(\theta_H)\mathrm{E}[N] - W_H(\theta_H)\big) \tag{2.8}$$

$$\text{s.t.} \quad W_L(\theta_L) - (c\theta_L + d_L(\theta_L))\,\mathrm{E}[N] \ge F, \tag{2.9}$$

$$W_H(\theta_H) - (c\theta_H + d_H(\theta_H))\,\mathrm{E}[N] \ge F, \tag{2.10}$$

$$W_L(\theta_L) - (c\theta_L + d_L(\theta_L))\,\mathrm{E}[N] \ge W_H(\theta_H) - (c\theta_H + d_L(\theta_H))\,\mathrm{E}[N], \tag{2.11}$$

$$W_H(\theta_H) - (c\theta_H + d_H(\theta_H))\,\mathrm{E}[N] \ge W_L(\theta_L) - (c\theta_L + d_H(\theta_L))\,\mathrm{E}[N]. \tag{2.12}$$

The incentive compatibility constraint (2.11) ensures that the low-risk health provider does not mimic the high-risk health provider. Likewise, the incentive compatibility constraint (2.12) ensures that the high-risk health provider does not mimic the low-risk health provider. To solve the problem (2.8)-(2.12), we first identify the relevant constraints, i.e., the constraints binding at the optimum of the health purchaser's problem. It is well-known from the mechanism design literature (for instance, see Mas-Colell et al. (1995)) that for the case of two agent types with one agent type having intrinsic tendency to mimic the other when faced with the first-best contract, only two constraints among the constraints (2.8)-(2.12) are binding: (1) the participation constraint of the type being mimicked, i.e., the constraint (2.9); and (2) the incentive compatibility constraint that prevent the type with mimicking tendency from mimicking the other, i.e., the constraint (2.12).

By maximizing the objective function (2.8) subject to the binding constraints (2.9) and (2.12), we can find the optimal effort level θ_L^* and θ_H^* and the associated expected reimbursement amounts $w_L^* = W_L(\theta_L^*)$ and $w_H^* = W_H(\theta_H^*)$. In Step 2, the health purchaser designs the set of contracts $\{w_{L,n}(\cdot)\}$ and $\{w_{H,n}(\cdot)\}$ such that the health providers find exerting the efforts θ_L^* and θ_H^* optimal.

Proposition 6 gives the second-best coordinating contracts when the health purchaser uses the gate-keeping contracts of Proposition 4.

Proposition 6 *When the health providers are located in low (L) and high (H) risk communities, under asymmetric information and verifiable number of customers, the gate keeping contracts*

$\{w^{SB}_{k,n}(z)\}$, *where* $w^{SB}_{k,n}(z) = w^{SB}_k z + b^{SB}_k(n-z)$, *for* $k \in \{L, H\}$, *coordinate the health purchaser-health provider relationships if:*

$$
\begin{aligned}
w^{SB}_L &= \frac{F}{\mathrm{E}[N]} + c + d'_L(\tilde{\theta}^{SB}_L) + d_L(\tilde{\theta}^{SB}_L) - \tilde{\theta}^{SB}_L d'_L(\tilde{\theta}^{SB}_L),\\
b^{SB}_L &= \frac{F}{\mathrm{E}[N]} + d_L(\tilde{\theta}^{SB}_L) - \tilde{\theta}^{SB}_L d'_L(\tilde{\theta}^{SB}_L),\\
w^{SB}_H &= \frac{F}{\mathrm{E}[N]} + c + d'_H(\tilde{\theta}^{SB}_H) + d_H(\tilde{\theta}^{SB}_H) - \tilde{\theta}^{SB}_H d'_H(\tilde{\theta}^{SB}_H) + d_L(\tilde{\theta}^{SB}_L) - d_H(\tilde{\theta}^{SB}_L),\\
b^{SB}_H &= \frac{F}{\mathrm{E}[N]} + d_H(\tilde{\theta}^{SB}_H) - \tilde{\theta}^{SB}_H d'_H(\tilde{\theta}^{SB}_H) + d_L(\tilde{\theta}^{SB}_L) - d_H(\tilde{\theta}^{SB}_L),
\end{aligned}
$$

where $\tilde{\theta}^{SB}_L$ *and* $\tilde{\theta}^{SB}_H$ *satisfy:*

$$\Pi'_L(\tilde{\theta}^{SB}_L) = c + d'_L(\tilde{\theta}^{SB}_L) + \frac{1-\alpha_L}{\alpha_L}\big(d'_L(\tilde{\theta}^{SB}_L) - d'_H(\tilde{\theta}^{SB}_L)\big), \tag{2.13}$$

$$\Pi'_H(\tilde{\theta}^{SB}_H) = c + d'_H(\tilde{\theta}^{SB}_H). \tag{2.14}$$

Proof. Conditions (2.13)-(2.14) follow from maximizing (2.8) subject to the binding constraints (2.9) and (2.12). The other conditions then follow from the health provider's optimality condition $w_k - b_k = c + d'(\tilde{\theta}_k)$, for $k \in \{L, H\}$ (see the proof of Proposition 4), and the binding constraints (2.9) and (2.12). □

Proposition 7 gives the second-best coordinating contracts when the health purchaser uses the piecewise linear contracts of Proposition 5.

Proposition 7 *When the health providers are located in low (L) and high (H) risk communities, under asymmetric information and verifiable number of customers, the sets of concave piecewise linear contracts* $\{w^{SB}_{L,n}(\cdot)\}$ *and* $\{w^{SB}_{H,n}(\cdot)\}$ *in which*

$$w^{SB}_{k,n}(z) = \begin{cases} b^-_{k,n} + w^-_{k,n} z, \ \textit{for } z \le z^0_{k,n}, \\ b^+_{k,n} + w^+_{k,n} z, \ \textit{for } z \ge z^0_{k,n}, \end{cases}$$

for $k \in \{L, H\}$, *can coordinate the health purchaser-health provider relationships if for any*

realization of N, say n:

$$
\begin{aligned}
z^0_{k,n} &= \tilde{\theta}^{SB}_k n, \ \textit{for } k \in \{L,H\},\\
w^-_{k,n} &= c + d'_k(\tilde{\theta}^{SB}_k) + \epsilon, \ \textit{for } k \in \{L,H\},\\
w^+_{k,n} &= c + d'_k(\tilde{\theta}^{SB}_k) - \epsilon, \ \textit{for } k \in \{L,H\},\\
b^-_{L,n} &= F + n\big(d_L(\tilde{\theta}^{SB}_L) - \tilde{\theta}^{SB}_L d'_L(\tilde{\theta}^{SB}_L)\big) - \tilde{\theta}^{SB}_L n\epsilon,\\
b^-_{H,n} &= F + n\big(d_H(\tilde{\theta}^{SB}_H) - \tilde{\theta}^{SB}_H d'_H(\tilde{\theta}^{SB}_H)\big) - \tilde{\theta}^{SB}_H n\epsilon + n\big(d_L(\tilde{\theta}^{SB}_L) - d_H(\tilde{\theta}^{SB}_L)\big),\\
b^+_{k,n} &= b^-_{k,n} + (w^-_{k,n} - w^+_{k,n})\tilde{\theta}^{SB}_k n, \ \textit{for } k \in \{L,H\},
\end{aligned}
$$

where $\tilde{\theta}^{SB}_L$ and $\tilde{\theta}^{SB}_H$ are calculated from Eq. 2.13 and Eq. 2.14, and $\epsilon > 0$ is large enought to make $w_{L,n}(\cdot)$'s and $w_{L,n}(\cdot)$'s concave for any $n \in [\underline{n}, \bar{n}]$.

Proof. Proof is similar to that of Proposition 6 and hence omitted. □

Corollary 1 *Under asymmetric information with verifiable number of customers, the optimal menu of gate-keeping or piecewise linear coordinating contracts entails:*

1. *No effort distortion for the high-risk health provider, i.e., $\tilde{\theta}^{SB}_H = \tilde{\theta}^{FB}_H$. A downward effort distortion for the low-risk health provider, i.e., $\tilde{\theta}^{SB}_L < \tilde{\theta}^{FB}_L$.*

2. *Only the high-risk health provider obtains a positive information rent equal to $\mathrm{E}[N]\big(d_L(\tilde{\theta}^{SB}_L) - d_H(\tilde{\theta}^{SB}_L)\big)$.*

Proof. See Appendix. □

Corollary 1 states that in order to achieve the welfare maximizing effort levels, the health purchaser should pay information rent to the high-risk health provider to induce him to exert the desired effort level. Also note that the information rent is proportional to the expected number of customers, and therefore, a high-risk health provider serving a larger population demands higher incentive to exert the desired effort level compared to a high-risk health providers serving a smaller population.

2.4 Coordinating Contracts with Unverifiable Number of Customers for the Medical Intervention

In the previous section, we defined "the number of customers for a medical intervention, N," as the number of individuals who visit the health provider to consume the medical intervention if prescribed by the health provider. We also assumed that any realization of N can be verified by the health purchaser, which is not necessarily the case in practice. For instance, consider the following two examples:

Example 1: Suppose that a health provider offers several different medical services. Consequently, to be able to verify the number of customers for a particular medical intervention, the health purchaser must elicit the purpose of each patient's visit to the health provider, and therefore, as the number of health providers and the population size increases, satisfying the verifiability assumption would become prohibitively costly or even impractical.

Example 2: Many preventive medical interventions are prescribed while the patient is visiting the health provider for other purposes. For instance, the health providers commonly take advantage of any occasion to persuade the patients at higher risk to undergone the appropriate cancer screening tests. Therefore, in these cases, "the number of customers for a particular medical intervention, N" is better defined as the number of individuals for whom the health provider has the *opportunity* of prescribing the intervention. For these situations, the number of customers for the intervention can hardly be verified even in the presence of an advanced information system.

In this section, we relax the verifiability assumption, and characterize the coordinating contracts for the health care systems where the number of customers for an underlying medical intervention is not verifiable by the health purchaser.

If the health provider chooses to exert the effort level θ, then for a realization of N, say n, the total number of people used the intervention equals $z = n\theta$. As a result, for an observed value of z, the health purchaser cannot determine the true value of the variables n and θ that resulted in the observed z. Hence, the health purchaser is faced with both the problem of asymmetric information (unverifiable n), and moral hazard (unverifiable θ).

2.4.1 Coordinating Contract with Homogenous Health Providers

Let $w(\cdot)$ denote a contract offered by the health purchaser to the health provider, that reimburses the health provider $w(z)$ dollars if the intervention is administered to z individuals. When the number of customers is not verifiable, the health purchaser solves the following optimization problem to find a contract $w(\cdot)$ that coordinates the health purchaser-health provider relationship:

$$\max_{\{w(\cdot)\},\theta} \quad \int_{\underline{n}}^{\bar{n}} \left(n\Pi(\theta) - w(\theta n)\right) f_N(n)dn \tag{2.15}$$

$$\text{s.t.} \quad \theta = \arg\max_{0\le\breve{\theta}\le 1} \int_{\underline{n}}^{\bar{n}} \left(w(\breve{\theta}n) - c\breve{\theta}n - nd(\breve{\theta})\right) f_N(n)dn \tag{2.16}$$

$$\int_{\underline{n}}^{\bar{n}} \left(w(\theta n) - c\theta n - nd(\theta)\right) f_N(n)dn \ge F. \tag{2.17}$$

Constraint (2.16) insures that under the contract $w(\cdot)$, the health provider's optimal effort choice is θ. The participation constraint (2.17) insures the health provider's interest in implementing the medical program.

The contract that is commonly used when the number of customers for the medical intervention, N, is not verifiable is the *risk-fee* or *fee-for-service* contract, defined below.

Definition 1 (Risk-free or fee-for-service contract) *A contract $w(\cdot)$ is called a risk-free contract if for a given $\tilde{\theta}$ and a realization N, say n, we have $w(\tilde{\theta}n) = c\tilde{\theta}n + nd(\tilde{\theta}) + F$, for any $n \in [\underline{n}, \bar{n}]$. By making the transformation $z = \tilde{\theta}n$, this contract will be of the bounded fee-for-service contract $w(z) = (c + \frac{d(\tilde{\theta})}{\tilde{\theta}})z + F$, for $\tilde{\theta}\underline{n} \leq z \leq \tilde{\theta}\bar{n}$.*

If the health purchaser offers a fee-for-service contract, a health provider facing with demand n solves the following problem to determine her effort level:

$$\max_{0\leq\theta\leq 1} (c + \frac{d(\tilde{\theta})}{\tilde{\theta}})\theta n - c\theta n - nd(\theta) - F. \tag{2.18}$$

The first derivative of the objective function (2.18) is $n(\frac{d(\tilde{\theta})}{\tilde{\theta}} - d'(\tilde{\theta}))$, which is always negative because of convexity of $d(\cdot)$. Therefore, the fee-for-service contract $w(z) = (c + \frac{d(\tilde{\theta})}{\tilde{\theta}})z + F$, $\tilde{\theta}\underline{n} \leq z \leq \tilde{\theta}\bar{n}$ leads the health provider to select an effort level which is less than the welfare-maximizing effort level $\tilde{\theta}$.

If we let the contract $w(\cdot)$ reimburse the health provider at rate $c + d'(\tilde{\theta})$ instead of $c + \frac{d(\tilde{\theta})}{\tilde{\theta}}$, then for any realization of demand, the health provider's profit does not change with his selected effort level, and the health provider will not have any incentive to deviate from $\tilde{\theta}$. Hence, when the health purchaser's contract entails a sufficiently large of contractual periods, it is reasonable to assume that if the effort level $\tilde{\theta}$ maximizes the health provider's expected profit over the entire contractual term, the health provider will exert the effort level $\tilde{\theta}$ in each period, irrespective of the realized demand. The following proposition characterizes a contract that coordinates the system under this condition.

Proposition 8 *When the number of customers for a medical intervention is not verifiable, and $\underline{n}f(\underline{n}) > 0$, $\bar{n}f(\bar{n}) > 0$, and $d'(\tilde{\theta}) < \frac{F + d(\tilde{\theta})\mathrm{E}[N] + c\tilde{\theta}\underline{n}}{(\mathrm{E}[N] - \underline{n})\tilde{\theta}}$, the bounded linear contract $w(\cdot)$ coordinates the health purchaser - health provider relationship if the number of contractual periods is sufficiently large and:*

$$w(z) = B + \big(c + d'(\tilde{\theta})\big)z, \text{ for } \tilde{\theta}\underline{n} \leq z \leq \tilde{\theta}\bar{n}, \tag{2.19}$$

where $B = F - \big(\tilde{\theta}d'(\tilde{\theta}) - d(\tilde{\theta})\big)\mathrm{E}[N]$ and $\tilde{\theta}$ is obtained from Eq. 2.2.

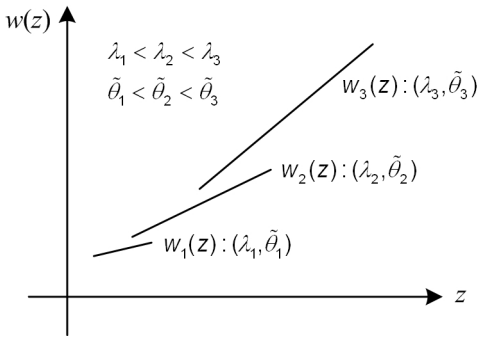


Figure 2.3: Coordinating Contract under Unverifiable Number of Customers

Proof. See Appendix. □

According to the contract of Proposition 8, at the beginning of the period, the health purchaser should make a transfer $B < F$ to the health provider, and then for each patient served, reimburses the health provider proportional to the service cost plus the *marginal* disutility at the optimal effort level, i.e., $c + d'(\tilde{\theta})$. Proposition 8 also specifies that a contract of simple form (2.19) can coordinate the health purchaser-health provider relationship only if $d'(\tilde{\theta})$ is smaller than a threshold. This condition is also intuitive, since if function $d'(\cdot)$ is too steep at point $\tilde{\theta}$, then the health provider will have the sufficient incentive to choose an effort level lower than $\tilde{\theta}$. Also note that the contract (2.19) does not depend on the shape of the distribution function $f_N(\cdot)$; however, for its successful implementation, the probability support $[\underline{n}, \bar{n}]$ should be agreed on by both the both the health purchaser and the health provider.

Figure 2.3 displays three contracts for different values of the health purchaser's WTP (λ). As λ increases the health purchaser is willing to offer a contract the implements the preventive intervention for a larger portion of the population (higher $\tilde{\theta}$).

The linear contract (2.19) possesses some restrictive properties as well. For instance, as $F \to 0$ and $\underline{n} \to 0$, the condition $d'(\tilde{\theta}) < \frac{F + d(\tilde{\theta})\mathrm{E}[N] + c\tilde{\theta}\underline{n}}{(\mathrm{E}[N] - \underline{n})\tilde{\theta}}$ becomes $\tilde{\theta} d'(\tilde{\theta}) < d(\tilde{\theta})$, which is never satisfied when $d(\cdot)$ is convex.

If the demand can be assumed to be a *discrete* random variable, then the health purchaser can circumvent the restriction posed by the contract (2.19) by designing a menu of incentive-feasible contracts. For any realization of demand, such a menu of contracts induces the health provider to exert the welfare-maximizing effort level, and then to reveal the true value of the realized demand.

Suppose that demand can be n_i with probability p_i, for $i \in \{1, \ldots, M\}$, $n_1 < n_2 < \ldots < n_M$, and $\sum_{i=1}^{M} p_i = 1$. The following sequence of events occurs: the health purchaser offers a set of incentive-feasible contacts $W = \{w_1(\cdot), \ldots, w_M(\cdot)\}$; the health provider decides whether to accept the contract or not; demand is realized, and the health provider exerts an effort level

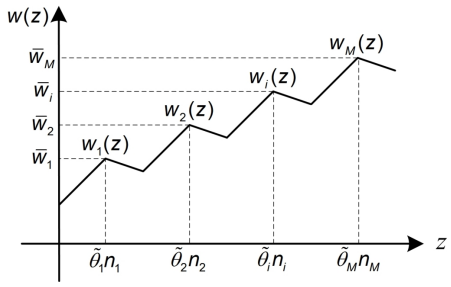


Figure 2.4: A menu of concanve piecewise linear contacts

that maximizes his profit, and then selects a contract in set W according to which he would like to be reimbursed. Note that by the Extended Revelation Principle (Laffont and Martimort 2001), when demand can take M values, the set W only needs to contain M contracts.

Figure 2.4 displays an example of a set of contacts $\{w_1(\cdot), \ldots, w_M(\cdot)\}$, in which $w_i(\cdot)$'s are assumed to be concave piece-wise linear contracts of form 2.7.

The health purchaser solves the following optimization problem to find the menu of incentive-feasible contracts $\{w_1(\cdot), \ldots, w_M(\cdot)\}$, along with the optimal effort levels $\tilde{\theta}_i$ to be exerted when demand is n_i, $i \in \{1, \ldots, M\}$.

$$\max_{\{w_i(\cdot)\}, 0 \le \theta_i \le 1} \sum_{i=1}^{M} p_i \big(n_i \Pi(\theta_i) - w_i(\theta_i n_i) \big) \tag{2.20}$$

$$\text{s.t. } \theta_i = \arg\max_{0 \le \theta \le 1} \{ w_i(\theta n_i) - c\theta n_i - n_i d(\theta) \}, \text{ for } i \in \{1, \ldots, M\}, \tag{2.21}$$

$$w_i(\theta_i n_i) - c\theta_i n_i - n_i d(\theta_i) \ge \max_{0 \le \theta \le 1} \{ w_j(\theta n_i) - c\theta n_i - n_i d(\theta) \}, \text{ for } j \ne i \in \{1, \ldots, M\}, \tag{2.22}$$

$$\sum_{i=1}^{M} \big(w_i(\theta_i n_i) - c\theta_i n_i - n_i d(\theta_i) \big) \ge F. \tag{2.23}$$

Constraint (2.21) specifies the optimal effort level exerted by the health provider when demand is n_i. Constraint (2.22) insures that a health provider with demand n_i selects to be reimbursed by contract $w_i(\cdot)$ and reveals the true demand. The participation constraint (2.23) insures that the health provider accepts the menu of contracts $\{w_1(\cdot), \ldots, w_M(\cdot)\}$. We use the piecewise linear contracts (2.7) to design the coordinating contracts, mainly because they are easy to characterize, and if they satisfy the constraints (2.21)-(2.22), they can be easily shifted to satisfy the participation constraint (2.23) in equality, which make these contracts optimal for the health purchaser.

To solve the optimization problem (2.20)-(2.23) for the piecewise linear contract (2.7), let's

assume that we know the effort level that maximizes the welfare for each demand n_i, i.e., $\tilde{\theta}_i$'s, and that the menu of contracts $\{w_1(\cdot), \ldots, w_M(\cdot)\}$ induces effort level $\tilde{\theta}_i$ for the health provider with demand n_i, and induces the effort level $\theta_j^* = \tilde{\theta}_i n_i / n_j$, for the health provider with demand n_j, who selects to be reimbursed according to $w_j(\cdot)$, $j \neq i$ (we will see later that these conditions can be easily satisfied).

By constraint (2.21) the health provider with demand n_i exerts effort level $\tilde{\theta}_i$, and selects to be reimbursed according to the contract $w_i(\cdot)$, which pays $w_i(\tilde{\theta}_i n_i)$ to the health provider. Let $\bar{w}_i = w_i(\tilde{\theta}_i n_i)$ (See Figure 2.4); now, the health purchaser's expected payment to the health provider will be equal to $\bar{W} = \sum_{i=1}^{M} p_i \bar{w}_i$. Hence, the expected health purchaser's payment (and the expected health provider's reimbursement) are independent of the shape of the contract in points other than $\{\tilde{\theta}_1 n_1, \ldots, \tilde{\theta}_M n_M\}$. Therefore, the health purchaser's problem reduces to finding the welfare-maximizing effort levels $\tilde{\theta}_i$'s and the reimbursements $\bar{w}_i$'s while assuring that the menu of contracts $\{w_1(\cdot), \ldots, w_M(\cdot)\}$ induces the effort level $\tilde{\theta}_i$ for the demand n_i.

Also, we assumed that the menu of contract $\{w_1(\cdot), \ldots, w_M(\cdot)\}$ induces the effort level $\theta_i^* = \tilde{\theta}_j n_j / n_i$ for a health provider with demand n_i who selects to be reimbursed according to $w_j(\cdot)$, for$j \neq i$; that is, if a health provider with demand n_i want to mimic a health provider with demand n_j for $j \neq i$, then his optimal effort level will be $\theta_i^* = \tilde{\theta}_j n_j / n_i$. This condition reduces the constraint (2.22) to:

$$w_i(\theta_i n_i) - c\theta_i n_i - n_i d(\theta_i) \geq w_j(\theta_j n_j) - c\theta_j n_j - n_i d(\frac{\theta_j n_j}{n_i}), \text{ for } j \neq i \in \{1, \ldots, M\}. \quad (2.24)$$

Lemma 3 *Constraint (2.24) is equivalent to:*

$$w_i(\theta_i n_i) - c\theta_i n_i - n_i d(\theta_i) \geq w_{i-1}(\theta_{i-1} n_{i-1}) - c\theta_{i-1} n_{i-1} - n_i d(\frac{\theta_{i-1} n_{i-1}}{n_i}), \text{ for } i \in \{2, \ldots, M\} \quad (2.25)$$

Proof. See Apendix. □

By Lemma 3, the health purchaser can now find the welfare-maximizing effort levels $\tilde{\theta}_i$'s , and the reimbursements $\bar{w}_i$'s by solving the following optimization problem:

$$\max_{\bar{w}_i, 0 \leq \theta_i \leq 1} \sum_{i=1}^{M} p_i \left(n_i \Pi(\theta_i) - \bar{w}_i\right) \quad (2.26)$$

$$\text{s.t. } \bar{w}_i - c\theta_i n_i - n_i d(\theta_i) \geq \bar{w}_{i-1} - c\theta_{i-1} n_{i-1} - n_i d(\frac{\theta_{i-1} n_{i-1}}{n_i}), \text{ for } i \in \{2, \ldots, M\} \quad (2.27)$$

$$\sum_{i=1}^{M} p_i \big(\bar{w}_i - c\theta_i n_i - n_i d(\theta_i)\big) = F. \quad (2.28)$$

Lemma 4 *The solution of the optimization problem (2.26)-(2.28) entails:*

1. *Unique optimal effort levels* $\tilde{\theta}_1 = \tilde{\theta}_2 = \cdots = \tilde{\theta}_M = \tilde{\theta}$, *where* $\tilde{\theta}$ *solves* $\Pi'(\tilde{\theta}) = c + d'(\tilde{\theta})$, *and*

2. *Optimal reimbursements* $\bar{w}_i^*$*'s lying on hyperplane* $\sum_{i=1}^{M} p_i \bar{w}_i^* = \big(c\tilde{\theta} + d(\tilde{\theta})\big) E[N] + F$.

Proof. See Apendix. □

According to Lemma 4, the optimal reimbursements $\bar{w}_i^*$'s are not unique (and characterized by a hyperplane); however, it is reasonable to determine $\bar{w}_i^*$'s as such to make the constraints (2.27) binding at optimality. Finally, the results in Lemma 3 and Lemma 4 rely on the assumption that the menu of contracts $\{w_1(\cdot), \ldots, w_M(\cdot)\}$ induces effort level $\tilde{\theta}_i$ for health provider with demand n_i, and induces the effort level $\theta_j^* = \tilde{\theta}_i n_i / n_j$ for health provider with demand n_j who selects to be reimbursed according to $w_j(\cdot)$, $j \neq i$. Therefore, we also need to determine the parameters of function $w_i(\cdot)$ as such to satisfy these assumptions. The following proposition summarizes the result assuming that $\bar{w}_i^*(\cdot)$'s satisfy the constrains (2.27) in equality.

Proposition 9 *When the number of customers for a medical intervention is not verifiable, the menu of piece-wise linear contracts* $\{w_1(\cdot), \ldots, w_M(\cdot)\}$, *in which* $w_i(\cdot)$*'s are of form (2.7), coordinates the health purchaser-health provider relationship if:*

$$z_i^0 = \tilde{\theta} n_i, \ \textit{for } i \in \{1, \ldots, M\}, \tag{2.29}$$

$$w_i^- = c + \max\{d'(\tilde{\theta}), \max_{1 \leq j \leq i-1} \{d'(\frac{\tilde{\theta} n_i}{n_j})\}\} + \epsilon, \ \textit{for } i \in \{1, \ldots, M\}, \tag{2.30}$$

$$w_i^+ = c + \min\{d'(\tilde{\theta}), \min_{i+1 \leq j \leq M} \{d'(\frac{\tilde{\theta} n_i}{n_j})\}\} - \epsilon, \ \textit{for } i \in \{1, \ldots, M\}, \tag{2.31}$$

$$b_i^- = \bar{w}_i - w_i^- \tilde{\theta} n_i, \ \textit{for } i \in \{1, \ldots, M\}, \tag{2.32}$$

$$b_i^+ = b_i^- + (w_i^- - w_i^+) \tilde{\theta} n_i, \ \textit{for } i \in \{1, \ldots, M\}, \tag{2.33}$$

$$\bar{w}_i = \begin{cases} h_1 + \sum_{i=2}^{M} \sum_{j=1}^{i-1} p_j h_i, \ i = M, \\ \bar{w}_M - \sum_{j=i+1}^{M} h_j, \ i = 1, \ldots, M-1, \end{cases} \quad \textit{for } i \in \{1, \ldots, M\}, \tag{2.34}$$

where $\tilde{\theta}$ *solves* $\Pi'(\tilde{\theta}) = c + d'(\tilde{\theta})$ *for* $i \in \{1, \ldots, M\}$, $h_1 = \big(c\tilde{\theta} + d(\tilde{\theta})\big) E[N] + F$, $h_i = c\tilde{\theta} n_i + n_i d(\tilde{\theta}) - c\tilde{\theta} n_{i-1} - n_i d(\frac{\tilde{\theta} n_{i-1}}{n_i})$, *for* $i \in \{2, \ldots, M\}$, *and* $\epsilon > 0$ *is large enought to make the contracts* $w_i(\cdot)$ *concave for* $i \in \{1, \ldots, M\}$.

Proof. See Apendix. □

2.4.2 Coordinating Contract with Heterogeneous Health Providers

In §2.4.1, we assumed that health providers are homogenous in terms of the patient populations. In this section, we consider two health providers, one in a high-risk population and the other in a low-risk population. Same as in §2.3.2, we assumed that the health purchaser is not able to identify the type of each health provider but she knows that a health provider is low-risk with probability α_L and high risk with probability α_H, where $\alpha_L + \alpha_H = 1$. To find the coordinating contracts, we follow the exact two steps described in §2.3.2. Proposition 10 and Proposition 11 summarize the results for bounded linear and set of piecewise linear contracts, respectively.

Proposition 10 *When the number of customers is not verifiable and the health providers are located in low (L) and high (H) risk communities, not observable by the health purchaser, the linear second-best coordinating contracts $w_L^{SB}(\cdot)$ and $w_H^{SB}(\cdot)$ are:*

1. $w_L^{SB}(z) = B_L + \big(c + d'_L(\tilde{\theta}_L^{SB})\big)z$, *for* $\tilde{\theta}_L^{SB}\underline{n} \le z \le \tilde{\theta}_L^{SB}\bar{n}$, *where* $B_L = F - \big(d'_L(\tilde{\theta}_L^{SB})\tilde{\theta}_L^{SB} - d_L(\tilde{\theta}_L^{SB})\big)\mathrm{E}[N]$ *and* $\tilde{\theta}_L^{SB}$ *is obtained from Eq. 2.13, provided that* $d'_L(\tilde{\theta}_L^{SB}) < \frac{F + d_L(\tilde{\theta}_L^{SB})\mathrm{E}[N] + c\tilde{\theta}_L^{SB}\underline{n}}{(\mathrm{E}[N] - \underline{n})\tilde{\theta}_L^{SB}}$, $\underline{n}f(\underline{n}) > 0$, *and* $\bar{n}f(\bar{n}) > 0$; *and*

2. $w_H^{SB}(z) = B_H + \big(c + d'_H(\tilde{\theta}_H^{SB})\big)z$, *for* $\tilde{\theta}_H^{SB}\underline{n} \le z \le \tilde{\theta}_H^{SB}\bar{n}$, *where* $B_H = F - \big(d'_H(\tilde{\theta}_H^{SB})\tilde{\theta}_H^{SB} - d_H(\tilde{\theta}_H^{SB})\big)\mathrm{E}[N] + \big(d_L(\tilde{\theta}_L^{SB}) - d_H(\tilde{\theta}_L^{SB})\big)\mathrm{E}[N]$ *and* $\tilde{\theta}_H^{SB}$ *is obtained from Eq. 2.14, provided that* $d'_H(\tilde{\theta}_H^{SB}) < \frac{F + d_H(\tilde{\theta}_H^{SB})\mathrm{E}[N] + c\tilde{\theta}_H^{SB}\underline{n}}{(\mathrm{E}[N] - \underline{n})\tilde{\theta}_H^{SB}}$, $\underline{n}f(\underline{n}) > 0$, *and* $\bar{n}f(\bar{n}) > 0$.

Proof. Proof is similar to that of Proposition 6, except that the health provider's optimality conditions are now given by Proposition 8. □

Proposition 11 *When the health providers are located in low (L) and high (H) risk communities, under asymmetric information and verifiable number of customers, the sets of concave piecewise linear contracts $\{w_{L,i}^{SB}(\cdot)\}$ and $\{w_{H,i}^{SB}(\cdot)\}$ in which*

$$w_{k,i}^{SB}(z) = \begin{cases} b_{k,i}^{-} + w_{k,i}^{-}z, \ \textit{for } z \le z_{k,i}^{0}, \\ b_{k,i}^{+} + w_{k,i}^{+}z, \ \textit{for } z \ge z_{k,i}^{0}, \end{cases}$$

for $k \in \{L, H\}$, can coordinate the health purchaser-health provider relationships if for any

realization of N, say n_i:

$$
\begin{aligned}
z^0_{k,i} &= \tilde{\theta}^{SB}_i n, \ \textit{for } k \in \{L,H\}, i \in \{1,\dots,M\},\\
w^-_{k,i} &= c + \max\{d'_k(\tilde{\theta}^{SB}_k), \max_{1\le j\le i-1}\{d'_k(\frac{\tilde{\theta}^{SB}_k n_i}{n_j})\}\} + \epsilon, \ \textit{for } k \in \{L,H\}, i \in \{1,\dots,M\},\\
w^+_{k,i} &= c + \min\{d'_k(\tilde{\theta}^{SB}_k), \min_{i+1\le j\le M}\{d'_k(\frac{\tilde{\theta}^{SB}_k n_i}{n_j})\}\} - \epsilon, \ \textit{for } k \in \{L,H\}, i \in \{1,\dots,M\},\\
b^-_{L,i} &= \bar{w}_{L,i} - w^-_{L,i}\tilde{\theta}^{SB}_L n_i, \ \textit{for } i \in \{1,\dots,M\},\\
b^-_{H,i} &= \bar{w}_{H,i} - w^-_{H,i}\tilde{\theta}^{SB}_H n_i + n_i\big(d_L(\tilde{\theta}^{SB}_L) - d_H(\tilde{\theta}^{SB}_L)\big), \ \textit{for } i \in \{1,\dots,M\},\\
b^+_{k,i} &= b^-_{k,i} + (w^-_{k,i} - w^+_{k,i})\tilde{\theta}^{SB}_k n_i, \ \textit{for } k \in \{L,H\}, i \in \{1,\dots,M\},
\end{aligned}
$$

$$
\bar{w}_{k,i} = \begin{cases} h_{k,1} + \sum_{i=2}^{M}\sum_{j=1}^{i-1} p_j h_i, \ i = M, \\ \bar{w}_{k,M} - \sum_{j=i+1}^{M} h_j, \ i = 1,\dots,M-1, \end{cases} \quad \textit{for } k \in \{L,H\}, i \in \{1,\dots,M\}, \tag{2.35}
$$

where $\tilde{\theta}^{SB}_L$ and $\tilde{\theta}^{SB}_H$ are calculated from Eq. 2.13 and Eq. 2.14, $h_{k,1} = \big(c\tilde{\theta}^{SB}_k + d(\tilde{\theta}^{SB}_k)\big)E[N] + F$, for $k \in \{L,H\}$, $h_{k,i} = c\tilde{\theta}^{SB}_k n_i + n_i d(\tilde{\theta}^{SB}_k) - c\tilde{\theta}^{SB}_k n_{i-1} - n_i d(\frac{\tilde{\theta}^{SB}_k n_{i-1}}{n_i})$, for $k \in \{L,H\}$ and $i \in \{2,\dots,M\}$, and $\epsilon > 0$ is large enought to make the contracts $w^{SB}_{k,i}(\cdot)$ concave for $k \in \{L,H\}$ and $i \in \{1,\dots,M\}$.

Proof. Proof is similar to that of Proposition 6, except that the health provider's optimality conditions are now given by Proposition 9. □

It can be also shown that Corollary 1 remains valid for the case of unverifiable number of customers.

2.5 Conclusions and Extensions

An immediate need for reforming health care payment systems has been stressed by medical scholars (Baron and Cassel 2008, Rosenthal 2008), which is mainly motivated by escalating health care costs, deviations from welfare maximizing medical resource allocations, and imbalance between primary and specialty care. Failing to consider the characteristics of the health providers (e.g., the provider's cost structure or the population's texture) is considered a major factor that makes a payment system result in health care inefficiency and inferior social welfare.

In this paper, we studied the contracts that coordinate the health purchaser-health provider relationship in a preventive health care delivery system. Such contracts allow the health purchaser and the health provider to optimize their objective functions while maximizing the population welfare. The proposed principal-agent model considers both the problem of moral

hazard (hidden action) and asymmetric information (hidden information). In this model, the health provider's decision about the rank of patients to whom the preventive intervention should be administered is not observable (hidden action). The health purchaser may also be unable to observe the number of customers for the medical intervention and the patient texture served by each health provider (hidden information).

When the number of customers for a medical intervention is verifiable by the health purchaser, we show that under certain conditions, there exist a gate-keeping contract and a set of concave piecewise linear contracts that coordinate the system. When the number of customers is not verifiable, we demonstrated that the widely-used fee-for-service payment system does not necessarily coordinate the health purchaser-health provider relationship, and the health provider tends to exert an effort level which is less than the optimum. Under these settings, the coordinating contract can be of bounded linear form that reimburses the health provider proportional to the unit service cost and the marginal disutility of the effort coordinates the health purchaser-health provider relationship, provided that the disutility at the optimal effort level is not too steep. A set of concave piecewise linear contracts can also coordinate the system under demand unverifiability when demand can be discretized.

Our model assumes that there is only one medical intervention available for the underlying disease. An immediate extension to our model will be finding the coordinating contracts for cases where more than one medical intervention is available for the disease. In the multiple-intervention paradigm, the coordinating contracts lead the health provider to prescribe the intervention for each patient that results in the best outcome.

Also, our model assumes that the health providers are faced with no capacity restriction in serving the patients. In reality, however, most of the health providers can only provide the intervention to a limited number of patients during the contractual period. Moreover, our model does not consider the possible competition among the health providers in attracting the patients. A model that captures the capacity restriction in a health care system or the competition among the health providers will be an interesting topic for future research.

Appendices

A.1. Proof of Proposition 4

The constant w and b in $w_n(z) = wz + b(n - z)$ should satisfy the rationality constraint (2.5) which results in $w - b = c + d'(\theta_n)$. Since w and b are constant, the health provider exerts the same optimal level for any realization of demand; i.e., $\theta_n = \theta^*$ for any n. Given this condition, the participation constraint (2.6) becomes $\mathrm{E}[N]\big((w - b)\theta^* + b - c\theta^* - d(\theta^*)\big)$. It is now easy to see that this participation constraint is binding at optimality, since otherwise, the health purchaser can increase her objective function by reducing w and b while still satisfying the health provider's optimality condition $w - b = c + d'(\theta^*)$. When the participation constraint (2.6) is binding, the objective function 2.4 becomes equivalent to maximizing the social welfare; hence, $\theta^* = \tilde{\theta}$. Finally, w and b can be determined through the health provider's optimality condition $w - b = c + d'(\tilde{\theta})$ and the binding constraint (2.6). □

A.2. Proof of Corollary 1

To show $\tilde{\theta}_H^{SB} = \tilde{\theta}_H^{FB}$, note that the first-best welfare maximizing effort level $\tilde{\theta}_H^{FB}$ solves $\Pi'_H(\tilde{\theta}_H^{FB}) = c + d'_H(\tilde{\theta}_H^{FB})$, and by Eq. 2.14, $\Pi'_H(\tilde{\theta}_H^{FB}) = c + d'_H(\tilde{\theta}_H^{FB})$. Comparing these two equations results in $\tilde{\theta}_H^{SB} = \tilde{\theta}_H^{FB}$. To show $\tilde{\theta}_L^{SB} < \tilde{\theta}_L^{FB}$, suppose the contrary: $\tilde{\theta}_L^{SB} \geq \tilde{\theta}_L^{FB}$. We know $\Pi'_L(\tilde{\theta}_L^{FB}) = c + d'_L(\tilde{\theta}_L^{FB})$. Since $d_L(\cdot)$ is strictly convex, $c + d'_L(\tilde{\theta}_L^{SB}) > c + d'_L(\tilde{\theta}_L^{FB})$. Therefore, by Eq. 2.13 and the assumption that $d'_L(\cdot) > d'_H(\cdot)$, we have $\Pi'_L(\tilde{\theta}_L^{SB}) > \Pi'_L(\tilde{\theta}_L^{FB})$; and since $\Pi'_L(\cdot)$ is strictly decreasing, it results in $\tilde{\theta}_L^{SB} < \tilde{\theta}_L^{FB}$, which contradicts the original assumption of $\tilde{\theta}_L^{SB} \geq \tilde{\theta}_L^{FB}$. The second part of the corollary is immediate from the contract parameters specified by Propositions 6 and 7. □

A.3. Proof of Proposition 8

To prove the proposition, we calculate the derivative of the health provider's benefit at the effort level $\tilde{\theta}$. As we will see later, the health provider's benefit is not differentiable at the point $\tilde{\theta}$, since the left-hand and the right-hand derivatives are different. Therefore, we calculate each separately. Let $\Pi_P(\theta)$ denote the health provider's benefit if he exert the effort θ, calculated by the objective function of problem (2.16). Since the participation condition (2.17) is binding at optimality, it implies that $\Pi_P(\tilde{\theta}) = 0$.

To calculate the left-hand derivate consider a point $\theta^1 < \tilde{\theta}$, sufficiently close to $\tilde{\theta}$. If the health provider chooses θ^1 instead of $\tilde{\theta}$, then for $n \in [\underline{n}, \frac{\tilde{\theta}n}{\theta^1}]$ he receives nothing from the health purchaser, and for $n \in [\frac{\tilde{\theta}n}{\theta^1}, \bar{n}]$ he will be reimbursed according to $w(z) = B + \big(c + d'(\tilde{\theta})\big)z$.

Therefore:

$$\begin{aligned}
\Pi_P(\tilde{\theta}) &- \Pi_P(\theta^1) \\
&= F - \int_{\underline{n}}^{\frac{\tilde{\theta}\underline{n}}{\theta^1}} \big(-c\theta^1 n - d(\theta^1)n\big) f(n)dn \\
&\quad - \int_{\frac{\tilde{\theta}\underline{n}}{\theta^1}}^{\bar{n}} \big(B + (c + d'(\tilde{\theta}))\theta^1 n - c\theta^1 n - nd(\theta^1)\big) f(n)dn \\
&= F + \big(c\theta^1 + d(\theta^1)\big) \int_{\underline{n}}^{\frac{\tilde{\theta}\underline{n}}{\theta^1}} nf(n)dn \\
&\quad - B\big(1 - F_N(\frac{\underline{n}\tilde{\theta}}{\theta^1})\big) - \big(d'(\tilde{\theta})\theta^1 - d(\theta^1)\big) \int_{\frac{\tilde{\theta}\underline{n}}{\theta^1}}^{\bar{n}} nf(n)dn \\
&= F - B + BF_N(\frac{\tilde{\theta}\underline{n}}{\theta^1}) - \big(d'(\tilde{\theta})\theta^1 - d(\theta^1)\big)\mathrm{E}[N] \\
&\quad + \big(c + d'(\tilde{\theta})\big)\theta^1 \int_{\underline{n}}^{\frac{\tilde{\theta}\underline{n}}{\theta^1}} nf(n)dn \qquad (2.36)
\end{aligned}$$

The left-hand derivative at point $\tilde{\theta}$ can be calculated by:

$$\begin{aligned}
\Pi_P'^{-}(\tilde{\theta}) &= \lim_{\theta^1 \to \tilde{\theta}} \frac{\Pi_P(\tilde{\theta}) - \Pi_P(\theta^1)}{\tilde{\theta} - \theta^1} \quad \text{(using L'Hopital's Rule)} \\
&= \lim_{\theta^1 \to \tilde{\theta}} B\frac{\tilde{\theta}\underline{n}}{(\theta^1)^2} f_N(\frac{\tilde{\theta}\underline{n}}{\theta^1}) + \big(d'(\tilde{\theta}) - d'(\theta^1)\big)\mathrm{E}[N] \\
&\quad (c + d'(\tilde{\theta}))\Big(\int_{\underline{n}}^{\frac{\tilde{\theta}\underline{n}}{\theta^1}} nf(n)dn + \theta^1 \frac{-\tilde{\theta}\underline{n}}{(\theta^1)^2} \frac{\tilde{\theta}\underline{n}}{\theta^1} f(\frac{\tilde{\theta}\underline{n}}{\theta^1})\Big) \\
&= \frac{\underline{n}f(\underline{n})}{\tilde{\theta}}\big(B + (c + d'(\tilde{\theta}))\tilde{\theta}\underline{n}\big) \\
&= \frac{\underline{n}f(\underline{n})}{\tilde{\theta}}\big(F + d(\tilde{\theta})\mathrm{E}[N] + c\tilde{\theta}\underline{n} + d'(\tilde{\theta})\tilde{\theta}(\underline{n} - \mathrm{E}[N])\big),
\end{aligned}$$

which is positive if $d'(\tilde{\theta}) < \frac{F + d(\tilde{\theta})\mathrm{E}[N] + c\tilde{\theta}\underline{n}}{(\mathrm{E}[N] - \underline{n})\tilde{\theta}}$. Hence, the health provider does not have incentive to choose an effort level less than $\tilde{\theta}$.

To calculate the right-hand derivate consider a point $\theta^1 > \tilde{\theta}$, sufficiently close to $\tilde{\theta}$. If the health provider now chooses θ^1 instead of $\tilde{\theta}$, then for $n \in [\underline{n}, \frac{\tilde{\theta}\bar{n}}{\theta^1}]$ he will be reimbursed according

to $w(z) = B + \big(c + d'(\tilde{\theta})\big)z$, and for $n \in [\frac{\tilde{\theta}\bar{n}}{\theta^1}, \bar{n}]$ he receives nothing. Therefore:

$$\begin{aligned}
\Pi_P(\theta^1) &- \Pi_P(\tilde{\theta}) \\
&= -F + \int_{\underline{n}}^{\frac{\tilde{\theta}\bar{n}}{\theta^1}} \big(B + \big(c + d'(\tilde{\theta})\big)\theta^1 n - c\theta^1 n - nd(\theta^1)\big) f(n)dn \\
&\quad + \int_{\frac{\tilde{\theta}\bar{n}}{\theta^1}}^{\bar{n}} \big(-c\theta^1 n - nd(\theta^1)\big)\, f(n)dn \\
&= -F + BF_N(\frac{\tilde{\theta}\bar{n}}{\theta^1}) + \big(d'(\tilde{\theta})\theta^1 - d(\theta^1)\big) \int_{\underline{n}}^{\frac{\tilde{\theta}\bar{n}}{\theta^1}} nf(n)dn \\
&\quad - \big(c\theta^1 + d(\theta^1)\big) \int_{\frac{\tilde{\theta}\bar{n}}{\theta^1}}^{\bar{n}} nf(n)dn \\
&= -F + BF_N(\frac{\tilde{\theta}\bar{n}}{\theta^1}) + \big(d'(\tilde{\theta})\theta^1 - d(\theta^1)\big)\mathrm{E}[N] \\
&\quad - \big(c + d'(\tilde{\theta})\big)\theta^1 \int_{\frac{\tilde{\theta}\bar{n}}{\theta^1}}^{\bar{n}} nf(n)dn
\end{aligned} \tag{2.37}$$

The right-hand derivative at point $\tilde{\theta}$ can be calculated by:

$$\begin{aligned}
\Pi_P'^{+}(\tilde{\theta}) &= \lim_{\theta^1 \to \tilde{\theta}} \frac{\Pi_P(\theta^1) - \Pi_P(\tilde{\theta})}{\theta^1 - \tilde{\theta}} \quad \text{(using L'Hopital's Rule)} \\
&= \lim_{\theta^1 \to \tilde{\theta}} -B\frac{\tilde{\theta}\bar{n}}{(\theta^1)^2} f_N(\frac{\tilde{\theta}\bar{n}}{\theta^1}) + \big(d'(\tilde{\theta}) - d'(\theta^1)\big)\mathrm{E}[N] \\
&\quad - \big(c + d'(\tilde{\theta})\big)\Big(\int_{\frac{\tilde{\theta}\bar{n}}{\theta^1}}^{\bar{n}} nf(n)dn - \theta^1 \frac{-\tilde{\theta}\underline{n}}{(\theta^1)^2}\frac{\tilde{\theta}\underline{n}}{\theta^1} f(\frac{\tilde{\theta}\underline{n}}{\theta^1})\Big) \\
&= -\frac{\bar{n}f(\bar{n})}{\tilde{\theta}}\big(B + c\tilde{\theta}\bar{n} + \bar{n}d(\tilde{\theta})\big) \\
&= -\frac{\bar{n}f(\bar{n})}{\tilde{\theta}}\big(F + d(\tilde{\theta})\mathrm{E}[N] + c\tilde{\theta}\bar{n} + d'(\tilde{\theta})\tilde{\theta}(\bar{n} - \mathrm{E}[N])\big),
\end{aligned}$$

which is always negative.

Also note that the contract $w(z) = \big(c + d'(\tilde{\theta})\big)z + B$ satisfies the constraint (2.17) in equality if $B = F - (d'(\tilde{\theta})\tilde{\theta} - d(\tilde{\theta}))\mathrm{E}[N]$. □

A.4. Proof of Lemma 3

Apparently, a health provider with demand n_i does not mimic a health provider with demand $n_j > n_i$ because of the convexity of disutility $d(\theta)$. Therefore, constraint (2.24) is redundant for all $j > i$. To establish the equivalency of (2.25) and (2.24) when $j < i$, we only need to show that for a given $i \geq 3$ if a health provider with demand n_i does not mimic a health provider

with demand n_{i-1}, and a health provider with demand n_{i-1} does not mimic a health provider with demand n_{i-2}, then the health provider with demand n_i does not mimic the health provider with demand n_{i-2} either. That is, if

$$w_i(\theta_i n_i) - c\theta_i n_i - n_i d(\theta_i) \geq w_{i-1}(\theta_{i-1} n_{i-1}) - c\theta_{i-1} n_{i-1} - n_i d(\frac{\theta_{i-1} n_{i-1}}{n_i}) \tag{2.38}$$

and

$$w_{i-1}(\theta_{i-1} n_{i-1}) - c\theta_{i-1} n_{i-1} - n_{i-1} d(\theta_{i-1}) \geq w_{i-2}(\theta_{i-2} n_{i-2}) - c\theta_{i-2} n_{i-2} - n_{i-1} d(\frac{\theta_{i-2} n_{i-2}}{n_{i-1}}), \tag{2.39}$$

then

$$w_i(\theta_i n_i) - c\theta_i n_i - n_i d(\theta_i) \geq w_{i-2}(\theta_{i-2} n_{i-2}) - c\theta_{i-2} n_{i-2} - n_i d(\frac{\theta_{i-2} n_{i-2}}{n_i}).$$

By convexity of $d(\cdot)$, $d(\alpha\theta) \leq \alpha d(\theta)$ for $0 \leq \alpha \leq 1$, which implies $d(\frac{n_{i-1}}{n_i}\theta_{i-1}) \leq \frac{n_{i-1}}{n_i} d(\theta_{i-1})$ using $\alpha = \frac{n_{i-1}}{n_i}$. Therefore, by (2.38),

$$\begin{aligned} w_i(\theta_i n_i) - c\theta_i n_i - n_i d(\theta_i) &\geq w_{i-1}(\theta_{i-1} n_{i-1}) - c\theta_{i-1} n_{i-1} - n_i d(\frac{\theta_{i-1} n_{i-1}}{n_i}) \\ &\geq w_{i-1}(\theta_{i-1} n_{i-1}) - c\theta_{i-1} n_{i-1} - n_{i-1} d(\theta_{i-1}) \\ &\geq w_{i-2}(\theta_{i-2} n_{i-2}) - c\theta_{i-2} n_{i-2} - n_i d(\frac{\theta_{i-2} n_{i-2}}{n_i}) \text{ (By Constraint (2.39)).} \end{aligned}$$

□

A.5. Proof of Lemma 4

Let $\mu_i \geq 0$ be the Lagrangian multipliers of constraints (2.27) and γ be the Lagrangian multiplier of constraint (2.28). The Lagrangian function of optimization problem (2.26)-(2.28) will be:

$$\begin{aligned} L = &\sum_{i=1}^{M} p_i\big(n_i \Pi(\theta_i) - \bar{w}_i\big) \\ &+ \sum_{i=1}^{M} \mu_i\big(\bar{w}_i - c\theta_i n_i - n_i d(\theta_i) - \bar{w}_{i-1} + c\theta_{i-1} n_{i-1} + n_i d(\frac{\theta_{i-1} n_{i-1}}{n_i})\big) \\ &+ \gamma\Big(\sum_{i=1}^{M} p_i\big(\bar{w}_i - c\theta_i n_i - n_i d(\theta_i)\big) - F\Big). \end{aligned} \tag{2.40}$$

The necessary optimality conditions for $\bar{w}_i$, $i \in \{1, \ldots, M\}$ are as follows:

$$\begin{aligned}
\frac{\partial L}{\partial \bar{w}_M} &= (\gamma - 1)\alpha_M + \mu_M = 0, \\
\frac{\partial L}{\partial \bar{w}_i} &= (\gamma - 1)\alpha_i + \mu_i - \mu_{i+1} = 0, \text{ for } i = \{2, \ldots, M-1\}, \\
\frac{\partial L}{\partial \bar{w}_1} &= (\gamma - 1)\alpha_1 - \mu_2 = 0.
\end{aligned}$$

Since by definition $\mu_M \geq 0$, $\gamma \geq 1$. Furthermore, it is apparent that at optimality, reducing F in constraint (2.28) by one unit increases the health purchaser's objective function by one unit; therefore, $\gamma^* = 1$, which results in $\mu_i^* = 0$ for all $i \in \{1, \ldots, M\}$. Thus, the optimality conditions for θ_i's imply that $\theta_1^* = \theta_2^* = \cdots = \theta_M^* = \theta^*$, where θ^* solves $\Pi'(\theta^*) = c + d'(\theta^*)$, and hence, $\theta^* = \tilde{\theta}$.

Also, since $\mu_i^* = 0$ for $i \in \{1, \ldots, M\}$, the constraints (2.27) can be either binding or non-binding at the optimality. In fact, any reimbursement scheme $(\bar{w}_1^*, \ldots, \bar{w}_M^*)$ that satisfies the constraint (2.28) is optimum; i.e., any reimbursement scheme $(\bar{w}_1^*, \ldots, \bar{w}_M^*)$ that lie on the hyperplane $\sum_{i=1}^M p_i \bar{w}_i^* = \big(c\theta^* + d(\theta^*)\big)\mathrm{E}[N] + F$.

To show that the first-order optimality conditions in Lemma 4 are also sufficient, note that setting $\gamma^* = 1$, and $\mu_i^* = 0$, $i \in \{1, \ldots, M\}$, reduce the Lagrangian function (2.40) to a strictly concave function in $(\theta_1, \ldots, \theta_M)$. □

A.6. Proof of Proposition 9

First we show that $\bar{w}_i$'s in (2.34) lies on the hyperplane of part (b) in Lemma 4:

$$\begin{aligned}
\sum_{i=1}^{M} p_i \bar{w}_i &= p_M \bar{w}_M + \sum_{i=1}^{M-1} p_i \big(\bar{w}_M - \sum_{j=i+1}^{M} h_j\big) = \bar{w}_M - \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} p_i h_j \\
&= h_1 + \sum_{i=2}^{M} \sum_{j=1}^{i-1} p_j h_i - \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} p_i h_j = h_1.
\end{aligned}$$

The last equality results from the fact that $\sum\limits_{i=2}^{M} \sum\limits_{j=1}^{i-1} p_j h_i = \sum\limits_{i=1}^{M-1} \sum\limits_{j=i+1}^{M} p_i h_j$. It is also straightforward to show that $\bar{w}_i$'s in (2.34) satisfy constraints (2.27) in equality.

Now we show that conditions (2.30) and (2.31) are sufficient for the menu of contracts $\{w_1(\cdot), \ldots, w_M(\cdot)\}$ to induce the effort level $\tilde{\theta}_i$ for health provider with demand n_i, and to induce the effort level $\theta_j^* = \tilde{\theta}_i n_i / n_j$, for health provider with demand n_j, $j \neq i$, who selects to be reimbursed according to $w_j(\cdot)$.

Apparently, for contract $w_i(\cdot)$ to induce the effort level $\tilde{\theta}_i$ when demand is n_i, it is sufficient

to have $w_i^- \geq c + d'(\tilde{\theta}_i) + \varepsilon$ and $w_i^+ \leq c + d'(\tilde{\theta}_i) - \varepsilon$, for $i = \{1, \ldots, M\}$.

Let θ_i^* denote the effort level exerted by a health provider with demand n_i who chooses the contract $w_j(\cdot)$, $j \geq i+1$. Then, two cases may occur:

Case 1: If $0 \leq \theta_i^* \leq \tilde{\theta}_j n_j / n_i$, then θ_i^* solves:

$$\max_{0 \leq \theta_i \leq \tilde{\theta}_j n_j / n_i} b_j^- + w_j^- \theta n_i - \theta n_i c - n_i d(\theta) \tag{2.41}$$

The first derivative of the objective function in (2.41) is $n_i\big(w_j^- - c - d'(\theta)\big)$, which is greater than $n_i\big(d'(\tilde{\theta}_j n_j/n_i) - d'(\theta)\big) > 0$ by (2.30) for $0 \leq \theta \leq \tilde{\theta}_j n_j / n_i$.

Case 2: If $\tilde{\theta}_j n_j / n_i \leq \theta_i^* \leq 1$, then θ_i^* solves:

$$\max_{\tilde{\theta}_j n_j / n_i \leq \theta_i \leq 1} b_j^+ + w_j^+ \theta n_i - \theta n_i c - n_i d(\theta) \tag{2.42}$$

The first derivative of the objective function in (2.42) is $n_i(w_j^+ - c - d'(\theta))$, which is less than $n_i(d'(\theta_j) - d'(\theta)) < 0$ by (2.31) for $\tilde{\theta}_j \leq \tilde{\theta}_j n_j / n_i \leq \theta \leq 1$ (note that $n_j \geq n_i$).

Therefore, for both cases $\theta_i^* = \tilde{\theta}_j n_j / n_i$. Likewise, we can show that by conditions (2.30) and (2.31), a health provider with demand n_i who chooses the contract $w_j(\cdot)$, $j \leq i-1$ always exerts effort level $\theta_i^* = \tilde{\theta}_j n_j / n_i$. The rest of the proof results from Lemma 4 and the structure of the contracts $w_i(\cdot)$'s given by (2.7). □

A.7. Healthcare Payment Systems

Over the past few decades, several reimbursement systems have been proposed and employed for health care systems. In what follows, we discuss each payment system briefly, and then review some recent work in this area. The reader is referred to Newhouse (1996) for a more thorough overview of the healthcare reimbursement systems and to McClellan (1997) for an empirical analysis.

One of the earliest reimbursement systems is the *cost-based* (*retrospective*) payment system in which the health provider is fully reimbursed for all costs of the medical services provided to the patient. Under cost-based reimbursement system, the health provider will make a minimum effort for cost reduction (Ma 1994). In a *fee-for-service* payment system, physicians receive a flat rate payment from the insurer for the service provided to the patient. Under this system, a profit-maximizing provider may tend to induce uninformed patients to consume more services than fully informed patients, or to sacrifice the quality in pursuit of achieving lower costs.

Similar to the fee-for-service payment system is the *prospective* payment system in which hospitals receive a payment dependent on the Diagnosis-Related Group (DRG) within which a patient falls. Prospective payment system essentially pays a fixed "price" per discharge, with

the price being determined by the patient's discharge diagnosis. Under uniform prospective payments for a heterogeneous population, the hospitals will have maximum incentive to avoid costly patients (Ma 1994).

In order to provide cost-reduction incentives to health providers, some health purchasers prefer a *capitated* payment system, also known as managed care, in which the health provider receives a flat payment per patient in a given time period and is then responsible for the provided service expenses. Nonetheless, similar to the prospective payment system, if the hospitals receive the same capitation for each patient, they have an incentive to avoid admitting high cost patients.

In an *outcome-adjusted* reimbursement, the patients receive treatment from the provider, who chooses the intensity of treatment and incurs an associated cost (Fuloria and Zenios 2001). The provider is then reimbursed by the purchaser according to observed patient outcomes. The purchaser's problem is to determine a payment system that induces the treatment choices maximizing the total societal welfare. This system can potentially make significant improvement in some certain healthcare systems; to implement, however, it requires accurate information about the treatment technology, patient characteristics, and the provider preferences (Fuloria and Zenios 2001).

The payment systems briefly described above along with their implications have been discussed in many studies over the past two decades. Table 2.1 summarizes a number of recent works with more details. In these studies, the interactions among the market entities are modeled in principal-agent frameworks, where a principal (e.g., health purchaser) delegates a task (e.g., providing medical service to the population) to an agent (e.g., health provider). The principal's problem is to design a contract that induces the agent to take the action(s) desired by the principal.

One stream of these studies focuses on the inefficiencies caused by *hidden information*, e.g., the purchaser has imperfect knowledge about the cost structure of the provider. These studies aim at finding the incentive-compatible contracts to reduce such inefficiencies. Another stream of research focuses on the *hidden action* problem, which occurs when the purchaser cannot observe the provider's action, for instance in treatment *effort* or *quality* of the delivered care. Quality means any aspect of service that benefits patients, whether during the process of treatment or in the health outcome after treatment (Chalkley and Malcomson 1998). Effort refers to any inputs by the physician that contribute to the intensity or quality of medical care but are difficult to measure or verify (Ma and McGuire 1997). Monitoring the quality of treatment and the physician's effort are either not possible or too costly to be worthwhile, and hence usually treated as hidden action.

Most of the healthcare reimbursement studies in Table 2.1 assume that only one medical intervention is available for an underlying disease. In Table 2.1, a homogeneous population

refers to population in which the individuals have the *same* risk of developing the disease or have the *same* prospective suffering from the disease, whereas in a heterogeneous population individuals are classified according to different *risk categories*. In a heterogeneous population, individuals with the same risk of contracting the disease and the same prospective devastation caused by the disease constitute a risk category.

Table 2.1: Literatures on health care payment systems

Reference	Agents Decision Variables	Contracts	Model	Intervention	Heterogeneity	Market Entities
Ma (1994)	Cost reduction effort, quality enhancement effort	Cost-based reimbursement, Prospective payment, Piecewise linear reimbursement	Hidden action	Single	Heterogeneous population	Health purchaser, Health provider
Ma and McGuire (1997)	Quality of care and treatment effort provided by the physician	Mixed of prospective and cost-based payment	Hidden action	Single	Homogeneous population	Health insurer, Physician, Patient
Chalkley and Malcomson (1998)	Number of patients treated, quality, effort for cost reduction	General form of capitation, cost-based, and prospective payment	Hidden action	Single	Heterogeneous population	Health purchaser, Health provider
Fuloria and Zenios (2001)	Intensity of treatment	Outcome-adjusted payment	Hidden action (Dynamic)	Single	Homogeneous population	Health purchaser, Health provider
Boadway et al. (2004)	Size of high-tech equipment, resources allocated to the hospital, patients assignment to equipments	Prospective payment, fee-for-service payment	Hidden information	Two (low tech, high tech)	Heterogeneous population	Health purchaser, Hospital, Physician
Jack (2005)	Effort, quality of care	Mixed of prospective and cost-based payment	Hidden information, Hidden action	Single	Heterogeneous health providers	Health purchaser, Health provider

Chapter 3

Coordinating Contracts in a Preventive Health Care System under Capacity Restrictions

Abstract

We consider a health care system consisting of two noncooperative parties: a health purchaser (e.g., a health insurer) and a health provider (e.g., a hospital). A principal-agent model is proposed to capture the interaction between the two parties. Offering a preventive intervention, the health provider should first decide how much medical capacity to allocate before observing the number of customers for the medical intervention. Then, having observed the demand, the health provider determines the type of patients who need to undergo the preventive medical intervention, and finally get reimbursed by the health purchaser based on the number of patients to whom the intervention is administered. We determine the contracts that *coordinate* the health purchaser-health provider relationship; i.e., the contracts that allow each entity to optimize its own objective function while maximizing the population's welfare. We show that (1) when the number of customers for the medical intervention is verifiable, a piecewise linear contact can coordinate the system; and (2) when the number of customers is not verifiable, a menu of incentive-feasible piecewise linear contacts can coordinate the system. We characterize the coordinating contracts under each setting.

Key words: coordinating contracts, health care, mechanism design, principal-agent models, capacity coordination.

3.1 Introduction

The health purchaser-health provider relationship is prevalent in health care systems. In this relationship, a health purchaser (e.g., health insurer) and a health provider (e.g., hospital) enter into a contractual agreement in which the health provider agrees to deliver service to the population it covers, and be reimbursed by the health purchaser according to a prespecified contract (payment system). A common example of such payment system is a *fee-for-service* contract, in which the health provider receives a flat rate for every complete episode of service provided to the patients.

The health purchaser-health provider relationship is plagued by incentive problems (Fuloria and Zenios 2001), especially since very often, the two parties have conflicting objectives. The health provider party possesses information advantage hidden from the health purchaser, or the actions taken by health provider are not observable and verifiable by the health purchaser. In this paper, we focus on a preventive health care system whose efficiency is highly dependent on the following two *hidden* actions: (1) the health provider's effort level in administering an underlying preventive intervention to the population, and (2) the level of medical capacity allocated by the health provider for offering the preventive intervention. We characterize the contracts that *coordinate* this relationship; i.e., the contracts allow both the health purchaser and health provider to optimize their own objective functions while maximizing the social welfare (Yaesoubi and Roberts 2009).

We consider a health care system consisting of a health purchaser, health providers, and a population. The population is assumed to be at risk of contracting or developing a disease for which one preventive medical intervention is available. The population is assumed to be *heterogeneous*, whose individuals are at different risk of developing the disease or prospective devastation caused by the disease. During each period, a random number of individuals visit the health provider to undergo the preventive medical intervention, if prescribed by the health provider. To each individual, we assign a number, called "rank," based on her expected level of devastation due to the disease, if once developed. An individual at higher risk of the disease is assigned a lower rank, and vice versa.

The health purchaser offers his contract to the health provider; assuming that the contract is acceptable, the health provider then allocates (or build) a certain amount of medical capacity, which is hidden from the health purchaser. The health provider then observes the demand for the medical intervention and depending on the availability of capacity, he specifies a risk *threshold* (hidden from the health purchaser), and administers the intervention to individuals of rank lower than the threshold. Finally, the health purchaser reimburses the health provider based on the prespecified contract.

This interaction may suffer significantly from the inefficiency caused by incentive problems.

First, the health provider may tend to invest in less capacity than the level desired from a societal perspective, since if demand for the medical intervention turns out to be low, he may face unutilized capacity. Second, once the capacity is allocated and the demand is realized, in order to maximize his profit, the health provider may have an incentive to choose a risk threshold that is different from the welfare maximizing choice.

Any verification and enforcement of capacity can be complex and costly, since capacity is affected by many factors such as facilities, equipments, nurses and physicians, scheduling, etc. Moreover, the health provider's risk-threshold may not be easily verified either. The only parameter that can be easily verified is the number of individuals who have undergone the intervention during the contractual period. The prevalent inefficiency caused by these incentive problems motivates us to pose a research question: What is the optimal payment system that reimburses the health provider based on the number of individuals who have undergone the intervention and *coordinate* the system? That is, the payment system that incents the health provider to allocate the level of capacity and to choose the risk-threshold as such to maximize the social welfare.

The remainder of paper is organized as follows: §3.2 details the model and defines the coordinating contracts given the underlying specifications. Section §3.3 reviews related literature in capacity and effort alignment. In §3.4, we characterize the coordinating contracts for cases where the number of customers for a medical intervention is verifiable by the health purchaser. We relax this condition in §3.5, and derive a menu of incentive feasible contracts that coordinates the system. Section 3.6 concludes the paper and discusses future expansions.

3.2 The Model

We consider a population which is at risk of developing or contracting a disease. The population can undergo a preventive intervention in order to reduce the risk of developing the disease or to alleviate the devastation caused by the disease once developed. We assume that the individuals can be ranked in a decreasing order of the expected devastation due to the disease, and a continuous value $\theta \in [0, 1]$ can be assigned to each individual such that $100\theta\%$ of population expect higher magnitude of suffering from the disease. The suffering from a disease may be the result of a financial loss or health deterioration. Financial loss is generally considered equal to the necessary direct and indirect treatment costs. To quantify health, we use Quality-Adjusted Life Year (QALY) which is an aggregate variable to represent both the length of life and the health quality of years lived (Miyamoto et al. 1998, Pliskin et al. 1980).

Let $q_j(\cdot)$ be a function returning the expected gain (or loss) in QALYs for a patient of rank θ who uses alternative $j \in \{0, 1\}$, where $j = 0$ and $j = 1$ denote 'not using' and 'using' the intervention, respectively. The function $q_j(\cdot)$ may include losses in QALY that might occur due

to the disease, normal aging, or an invasive intervention. Let $v_j(\cdot)$ be a function returning the expected treatment costs for a patient of rank θ who uses alternative $j \in \{0,1\}$, excluding the price of the medical intervention obtained.

Let λ denote the health purchaser's willingness-to-pay (WTP) for health, and define the function $\pi_j(\cdot)$ as $\pi_j(\theta) = v_j(\theta) + \lambda q_j(\theta)$ for $j \in \{0,1\}$. The function $\pi_j(\cdot)$ returns the health purchaser's net monetary benefit if the patient of rank θ uses the alternative$j \in \{0,1\}$. Define $\Pi(\theta) = \int_0^\theta (\pi_1(t) - \pi_0(t))\, dt$ as the health purchaser's gain if individuals of rank θ and less undergo the intervention. We assume that the functions $v_j(\theta)$ and $q_j(\theta)$, for $j \in \{0,1\}$, satisfy the assumptions 1 and 2 in (Yaesoubi and Roberts 2009) such that the function $\pi_1(\theta) - \pi_0(\theta)$ is decreasing in θ; and therefore, $\Pi'(\cdot) > 0$ and $\Pi''(\cdot) < 0$.

To offer the intervention, the health provider should allocate (or build) sufficient amount of medical capacity. Without loss of generality, we assume that to be able to administer the intervention to x individuals, x units of capacity is required. The health provider incurs a fixed cost $K(x)$ for allocating x unit of capacity. The fixed cost $K(x)$ includes the cost of the required medical resources and the cost of lost opportunities incurred by the health provider for allocating x unit of capacity. The fixed cost $K(\cdot)$ is assumed to be monotonically increasing and convex in x. Provided that adequate capacity is available, the health provider also incurs a variable cost c for each patient who uses the intervention, and a cost (or disutility) $d(\theta)$ for administering the intervention to the individuals of rank less than or equal to θ. The disutility $d(\cdot)$ is incurred mainly because as an individual's risk of contracting the disease reduces, it takes more effort to persuade her to use the intervention. The disutility $d(\cdot)$ is assumed to be monotonically increasing and convex in θ.

Let the random variable N denote the number of customers for the medical intervention who visit the health provider during the contractual period. To be more precise, we define "the number of customers for a medical intervention, N," as the number of individuals who visit the health provider to consume the medical intervention if prescribed by the health provider. For instance, during an epidemic period, N would be the number of individuals who consider themselves at risk of contracting the disease and visit the health provider in order to seek the preventive procedure. We assume that N has probability density $f_N(\cdot)$ and support $[\underline{n}, \bar{n}]$.

Now suppose that the health provider allocates x unit of medical capacity and administers the intervention to individual with rank lower than θ. Therefore for a given realization of N, say n, the intervention will be administered to θn patients. Provider that enough medical capacity is available to serve the θn patients, we assume that the social welfare can be represented by:

$$\text{Social Welfare} = (\Pi(\theta) - c\theta - d(\theta))\, \mathrm{E}[N] - K(x). \tag{3.1}$$

If we assume that the health purchaser is social welfare maximizer, then the health purchaser's problem is to find the medical capacity to allocate ($\tilde{x}$), and the splitting point $\tilde{\theta}$ such

that the individuals of rank $\theta \in [0, \tilde{\theta}]$ use the intervention and the individuals of rank $\theta \in [\tilde{\theta}, 1]$ do not use the intervention, while the social welfare (3.1) is maximized. Greater $\tilde{\theta}$ implies that a larger percentage of population should undergo the intervention, which requires greater implementation *effort*. Therefore, to conform to the contract theory literature, we refer to θ as the *effort* to be exerted by the health provider in implementing the intervention.

The main difficulty in optimizing the social welfare (3.1) is the fact that the decision variables x and θ are determined in different stages; that is, the capacity decision is generally made before observing the demand (n), and the threshold θ is usually determined after observing the demand based on the available capacity. Therefore, x is a first-stage decision variable and θ is a second-stage decision variable which depends on the realization of the random variable N.

To maximize the social welfare (3.1), we solve the following two-stage stochastic program:

$$\max_{x \geq 0} \Omega(x) = -K(x) + \mathrm{E}_N[Q(x, n)], \tag{3.2}$$

where

$$Q(x, n) = \max\{n\Pi(\theta) - c\theta n - nd(\theta) | \theta n \leq x, 0 \leq \theta \leq 1\} \tag{3.3}$$

is the recourse function. The recourse function $Q(x, n)$ has the following solution:

$$\theta^* = \begin{cases} \theta^0, \text{ if } \theta^0 n \leq x, \\ x/n, \text{ if } \theta^0 n \geq x, \end{cases}$$

where θ^0 solves $\Pi'(\theta^0) = c + d'(\theta^0)$. Now problem (3.2) becomes equivalent to:

$$\begin{aligned} \max_{x \geq \theta^0 \underline{n}} \Omega(x) = &- K(x) + \int_{\underline{n}}^{x/\theta^0} \left(n\Pi(\theta^0) - cn\theta^0 - nd(\theta^0)\right) f_N(n) dn \\ &+ \int_{x/\theta^0}^{\bar{n}} \left(n\Pi(\frac{x}{n}) - cx - nd(\frac{x}{n})\right) f_N(n) dn. \end{aligned} \tag{3.4}$$

Proposition 12 *To maximize the social welfare (3.1), the (unique) optimal capacity level $\tilde{x}$ satisfies:*

$$\int_{\tilde{x}/\tilde{\theta}}^{\bar{n}} \left(\Pi'(\frac{\tilde{x}}{n}) - d'(\frac{\tilde{x}}{n})\right) f_N(n) dn = c(1 - F(\frac{\tilde{x}}{\tilde{\theta}})) + K'(\tilde{x}), \tag{3.5}$$

and, for any realization of demand in the second stage, say n, the optimal threshold $\tilde{\theta}(n)$ satisfies,

$$\tilde{\theta}(n) = \begin{cases} \tilde{\theta}, \text{ if } \tilde{\theta} n \leq \tilde{x}, \\ \tilde{x}/n, \text{ if } \tilde{\theta} n \geq \tilde{x}, \end{cases} \tag{3.6}$$

where $\tilde{\theta}$, called welfare maximizing threshold, solves $\Pi'(\tilde{\theta}) = c + d'(\tilde{\theta})$.

Proof. See Appendix. □

Note that to find the social welfare maximizing capacity level $\tilde{x}$ by Eq. 3.5, simple line search can be employed since the left (right)-hand side of Eq. 3.5 is strictly decreasing (increasing) in $\tilde{x}$.

Proposition 12 assumes that the health provider's effort level, $\tilde{\theta}$, and the allocated capacity, $\tilde{x}$, is verifiable and that the health purchaser can observed the rank of each individual who visits the health provider. In practice, however, it is only the health provider who can observe the true rank of each patient and make the recommendations accordingly. It would be too expensive for the health purchaser to obtain the true value of all the patients' rank (θ's) for whom the intervention has been prescribed. Moreover, as mentioned before, any verification and enforcement of capacity can be very complex. Therefore, the health purchaser's problem is to design a contract that leads the health provider to choose the welfare maximizing effort level $(\tilde{\theta})$)and the level of capacity $(\tilde{x})$ as his optimal strategy. We say that a contract *coordinates* the health purchaser-health provider relationship if it implements the welfare maximizing effort level, $\tilde{\theta}$, and welfare maximizing capacity level, $\tilde{x}$, while allowing the health purchaser and health provider to optimize their objective functions.

The following sequence of events occurs: the health purchaser announces the contract; had found the contract agreeable, the health provider allocates medical capacity. Demand is realized, and the health provider determines his effort level based on the observed demand and the available capacity. Finally, the health purchaser reimburses the health provider according to the contract.

Let $w(\cdot)$ denote a contract offered by the health purchaser to the health provider that reimburses the health provider $w(z)$ dollars if the health provider administers the preventive intervention to z individuals. To maximize her profit, the health provider solves the following two-stage stochastic program:

$$\max_{x \geq 0} \Omega_R(x) = -K(x) + \mathrm{E}_N[Q_R(x, n)], \tag{3.7}$$

where

$$Q_R(x, n) = \max\{w(\theta n) - c\theta n - nd(\theta) | \theta n \leq x, 0 \leq \theta \leq 1\} \tag{3.8}$$

is the recourse function. Let x^* denote the solution of problem (3.7). We assume that the health provider finds the contract $w(\cdot)$ acceptable only if the following normalized *participation constraint* is satisfied:

$$-K(x^*) + \mathrm{E}_N[Q_R(x^*, n)] \geq 0. \tag{3.9}$$

The health purchaser now solves the following problem to find the contract $w(\cdot)$:

$$\max_{\substack{x\geq 0, 0\leq\theta\leq 1,\\ w(.)}} \quad \Omega_H(x) = \mathrm{E}_N[n\Pi(\theta) - w(n\theta)] \tag{3.10}$$

$$\text{s.t.} \quad x = \arg\max_{\breve{x}\geq 0}\{-K(\breve{x}) + \mathrm{E}_N[Q_R(\breve{x}, n)]\}, \tag{3.11}$$

$$\theta = \arg\max_{0\leq\breve{\theta}\leq 1}\{w(\breve{\theta}n) - c\breve{\theta}n - nd(\breve{\theta})|\breve{\theta}n \leq x, 0 \leq \breve{\theta} \leq 1\}, \forall n \in [\underline{n}, \bar{n}], \tag{3.12}$$

$$-K(x) + \mathrm{E}_N[Q_R(x, n)] \geq 0. \tag{3.13}$$

The *rationality constraints* (3.11) and (3.12) assure that the health provider allocate the level of capacity and effort desired by the health purchaser. The participation constraint (3.13) assures the acceptance of contract $w(\cdot)$ by the health provider. To solve the problem (3.10)-(3.13), we consider contracts with following property: if the contract $w(\cdot)$ satisfies the rationality constraints (3.11) and (3.12), then the contract $w(\cdot) + b$, where b is a constant, also satisfy the constraints (3.11) and (3.12). By this property, it is apparent that the participation constraint (3.13) is binding at the optimum, since otherwise, the health purchaser can shift down the contract $w(\cdot)$ to increase his objective function without violating the rationality constraints (3.11) and (3.12).

In finding the coordinating contracts, we suppose that the following assumption from (Yaesoubi and Roberts 2009) holds about the health provider.

Assumption 5 *The health provider is altruistic; that is, if for a given contract, exerting the welfare maximizing effort level $\tilde{\theta}$ is optimal for the health provider, then he administers the intervention to an individual of rank θ if and only if $\theta \leq \tilde{\theta}$.*

3.3 Literature Review

Several reimbursement systems have been proposed and employed for health care systems. The reader is referred to Newhouse (1996) for a thorough overview of the health care reimbursement systems and to the online companion attached to (Yaesoubi and Roberts 2009) for a brief review on some recent work in this area. In these studies, the interactions between the health purchaser and health provider is modeled in principal-agent frameworks, where a principal (e.g., health purchaser) delegates a task (e.g., providing medical service to the population) to an agent (e.g., health provider). The principal's problem is to design a contract that induces the agent to take the action(s) desired by the principal.

One stream of these studies focuses on the inefficiencies caused by *hidden information*, e.g., the purchaser has imperfect knowledge about the cost structure of the provider. These

studies aim at finding the incentive-compatible contracts to reduce such inefficiencies (Boadway et al. 2004, Jack 2005, Shleifer 1985). Another stream of research focuses on the *hidden action* problem, which occurs when the purchaser cannot observe the provider's action, such as, the *intensity* of treatment (Fuloria and Zenios 2001, Ma and McGuire 1997) or the *quality* of the delivered care (Jack 2005, Chalkley and Malcomson 1998, Ma 1994). Monitoring the quality of treatment and the physician's effort are either not possible or too costly to be worthwhile, and hence usually treated as hidden action.

When the main objective is to optimize the *global* system, the problem of moral hazard and asymmetric information can be accommodated through "coordinating contracts." Coordinating contracts have gained significant attention in the operations management literature (for a comprehensive review, refer to Cachon (2003)), and different types of coordinating contracts have been introduced and studied; for instance, see wholesale price contracts (Cachon 2003), payback contracts (Pasternack 1985), revenue-sharing contracts (Cachon and Lariviere 2005), and quantity-discount contract (Weng 1995).

Recently, Yaesoubi and Roberts (2009) extended the concept of the coordinating contracts to a preventive health care system consisting of a health purchaser and a health provider. In their model, the health provider administers the preventive intervention to the individual with rank less than a threshold (which is hidden from the health purchaser) and then gets reimbursed by the health purchaser based on the number of patients who received the preventive intervention during the contractual period. Their model holds the assumption that the health provider is not restricted by capacity in rendering service, and can administer the intervention to as many individual as desired during the contractual period. In many preventive health systems, however, limited capacity can be an important barrier in sustaining the desired level of service that maximizes the social welfare.

Capacity alignment has been extensively studied in the operations management literature as well. These studies, however, have mainly focused on the relationship between a manufacturer and a supplier in which the overall production is bounded by the supplier's capacity, or on a retailer and a manufacturer relationship in which the overall sale is bounded by the manufacturer's capacity. In the supplier-manufacturer problem, the manufacturer faces uncertain demand with a known prior distribution. The manufacturer offers the supplier a contract along with his initial demand forecast to build capacity. Assuming the contract is acceptable, the supplier then builds capacity. The manufacturer observes the demand and submits a final order. Finally, the supplier produces as much of the order as the capacity allows. Thus, the manufacturer's problem here is to induce the supplier to build enough capacity to optimize the supply chain.

This problem has been studied in different settings. Wang and Gerchak (2003) study a capacity investment game between an assembler and its component suppliers. Tomlin (2003)

studies the capacity investment game between a manufacturer and a supplier and introduces share-the-gain contracts in which the parties share the gain of high demand rather than the pain of low demand. Bernstein and DeCroix (2004) study multi-tier decentralized assembly system and characterize the optimal pricing and capacity choices in equilibrium. Chakravarty and Zhang (2007) focus on the incentive issues in a lateral relationship between two firms collaborating on capacity investment with information asymmetry. Ozer and Wei (2006) study the problem of optimal capacity decision under asymmetric forecast information, and structure contracts that induce credible forecast information sharing.

The characteristics of the model presented in §3.2 departs from the assumptions of these studies on supply chain capacity alignments in two aspects: (1) in the context supply chain, it is generally assumed that the marginal cost of capacity is constant; and (2) once the capacity is built, to optimize the supply chain, the capacity should be fully utilized (since the unit selling price is assumed to be higher than the unit production cost). In preventive health care systems, however, these assumptions do not necessarily hold. First, health providers (specifically hospitals and medical centers) offer multiple services while operating under limited medical resources and facility; therefore, the marginal cost of capacity allocated for a particular intervention will be monotonically increasing. Second, even if sufficient capacity is available, maximizing the social welfare may require *partial* consumption of the capacity.

3.4 Coordinating Contract with Verifiable Number of Customers for the Medical Intervention

In this section, we assume that the number of customers for the underlying medical intervention who visit the health provider, N, is verifiable by the health purchaser. For instance, consider a temporary health care facility offering a single preventive intervention (such as a vaccine) during an epidemic period; the individuals who consider themselves at risk of contracting the disease, visit the health provider, and the health provider administer the vaccine only to those with rank lower than the prespecified threshold. In such systems, the health purchaser can easily verify the number of customers for the preventive intervention by counting the total number of visits to the health provider. This assumption may also be valid when the health purchaser benefits from an advanced information system through which he can distinguish among the customers for different medical services. Nonetheless, as discussed in §3.5, satisfying this assumption can become very expensive or in some cases impossible. In this section, however, we suppose that this assumption holds.

Throughout this section, the subscript n in the contract $w_n(\cdot)$ is to emphasize the verifiability of the number of customers, and the fact that the reimbursement $w_n(z)$ depends on both the realized number of customers, n, and the number of patients who used the intervention, z.

When the number of customers is verifiable, we denote the health purchaser's contract by $\{w_n(\cdot)\}$, since it is in fact a collection of several functions $w_n(\cdot)$, each of which is designed for a particular realized number of customers, n.

3.4.1 Gate-Keeping Contracts

Consider a contract by which the health provider receives a payment w for each individual to whom the intervention is administered, and a payment b for each individual who does not use the intervention. Therefore, this contract will have the form of $\{w_n^G(z) = wz + b(n-z)\}$. For this contract, the recourse problem (3.8) will be equivalent to:

$$Q_R(x,n) = \max\{w\theta n + b(n-\theta n) - c\theta n - nd(\theta) | \theta n \leq x, 0 \leq \theta \leq 1\}. \tag{3.14}$$

For a given realization of demand n and capacity level x, the solution of recourse problem $Q_R(x,n)$ in (3.14) is:

$$\theta^*(n) = \begin{cases} \theta^0, \text{ if } \theta^0 n \leq x, \\ x/n, \text{ if } \theta^0 n > x, \end{cases} \tag{3.15}$$

where θ^0 satisfies

$$w - b = c + d'(\theta^0). \tag{3.16}$$

If the health provider decides to exert effort level θ^0 in the second stage, then she solves the following problem to determine the optimal level of capacity (substitute $w(\cdot)$ with $\{w_n^G(z) = wz + b(n-z)\}$ in problem (3.7)):

$$\begin{aligned} \max_{x \geq \theta^0 \underline{n}} \Omega_R(x) = &- K(x) + \int_{\underline{n}}^{x/\theta^0} \left((w-b-c)\theta^0 + b - d(\theta^0)\right) n f_N(n) dn \\ &+ \int_{x/\theta^0}^{\bar{n}} \left((w-b-c)x - nd(\frac{x}{n})\right) f_N(n) dn. \end{aligned} \tag{3.17}$$

It is easy to show that the sufficient condition for x^0 to solve the problem (3.17) is:

$$w - b = \frac{\int_{x^0/\theta^0}^{\bar{n}} d'(x^0/n) f_N(n) dn + K'(x^0)}{1 - F_N(x^0)} + c. \tag{3.18}$$

Also, the health purchaser determines w and b such that the participation constraint (3.9) becomes binding at the health provider's choice of x and θ. Therefore, by setting the objective

function $\Omega(x)$ in (3.17) equal to zero we get another relationship between w and b:

$$\begin{aligned}&\left(\theta^0\mathrm{E}[N]+\int_{x^0/\theta^0}^{\bar{n}}(x^0-\theta^0 n)f_N(n)dn\right)w+\left((1-\theta^0)\mathrm{E}[N]-\int_{x^0/\theta^0}^{\bar{n}}(x^0-\theta^0 n)f_N(n)dn\right)b\\&=\left(c\theta^0+d(\theta^0)\right)\mathrm{E}[N]+\int_{x^0/\theta^0}^{\bar{n}}\left(c(x^0-\theta^0 n)+n(d(\tfrac{x^0}{n})-d(\tilde{\theta}))\right)f_N(n)dn+K(x^0).\end{aligned}\tag{3.19}$$

The following proposition shows the conditions under which the gate-keeping contract coordinates the system.

Proposition 13 *When the number of customers is verifiable, the gate-keeping contract* $\{w_n^G(z)=wz+b(n-z)\}$ *coordinates the system if and only if:*

$$d'(\tilde{x})=\frac{\int_{\tilde{x}/\tilde{\theta}}^{\bar{n}}d'(\tilde{x}/n)f_N(n)dn+K'(\tilde{x})}{1-F_N(\tilde{x})},\tag{3.20}$$

and w *and* b *satisfy Eq. 3.16 and Eq. 3.19 for* $\theta^0=\tilde{\theta}$ *and* $x^0=\tilde{x}$.

If condition (3.20) is not satisfied, the gate-keeping contract can only coordinate either effort level or capacity, and not both. To coordinate the effort level, w *and* b *should satisfy Eq. 3.16 and Eq. 3.19, and to coordinate the capacity,* w *and* b *should satisfy Eq. 3.18 and Eq. 3.19, for* $\theta^0=\tilde{\theta}$ *and* $x^0=\tilde{x}$.

3.4.2 Linear (Fee-For-Service) Contracts

A linear (fee-for-service) contract $\{w_n^L(\cdot)\}$ is of the form $\{w_n^L(z)=F+wz\}$, where z is the number of individuals to whom the intervention is administered during the contractual period, $F\geq 0$ is a fixed transfer payment made at the beginning of the contractual period, and $w>c$ is a constant service price. Note that in many systems, the transfer payment F is set to zero.

Under the linear contract $\{w_n^L(\cdot)=F+cz\}$, for a given realization of demand n and capacity level x, the solution of recourse problem $Q_R(x,n)$ in (3.8) is:

$$\theta^*(n)=\begin{cases}\theta^0, \text{ if } \theta^0 n\leq x,\\ x/n, \text{ if } \theta^0 n\geq x,\end{cases}\tag{3.21}$$

where θ^0 satisfies

$$w=c+d'(\theta^0).\tag{3.22}$$

To find the capacity level in the first stage, the health provider solves the following problem:

$$\max_{x \geq \theta^0 \underline{n}} \Omega_R(x) = -K(x) + \int_{\underline{n}}^{x/\theta^0} (wn\theta^0 - cn\theta^0 - nd(\theta^0)) f_N(n) dn + \int_{x/\theta^0}^{\bar{n}} (wx - cx - nd(\frac{x}{n})) f_N(n) dn + F. \tag{3.23}$$

It is easy to show (see the proof of Proposition 14) that the sufficient condition for x^0 to solve the problem (3.23) is:

$$w = \frac{\int_{x^0/\theta^0}^{\bar{n}} d'(x^0/n) f_N(n) dn + K'(x^0)}{1 - F_N(x^0)} + c. \tag{3.24}$$

Also, the health purchaser determines w and F such that the participation constraint (3.9) becomes binding at the health provider's choice of x and θ. Therefore, by setting the objective function $\Omega_R(x)$ in (3.23) equal to zero we get another relationship between w and F:

$$F = K(x^0) + \left(w\theta^0 - c\theta^0 - d(\theta^0)\right) \mathrm{E}[N] - \int_{x^0/\theta^0}^{\bar{n}} \left((w-c)(x^0 - \theta^0 n) - n(d(\frac{x^0}{n}) - d(\theta^0))\right) f_N(n) dn. \tag{3.25}$$

Proposition 14 *When the number of customers is verifiable, the linear contract* $\{w_n^L(z) = F + wz\}$ *coordinates the system if and only if:*

$$d'(\tilde{x}) = \frac{\int_{\tilde{x}/\theta^0}^{\bar{n}} d'(\tilde{x}/n) f_N(n) dn + K'(\tilde{x})}{1 - F_N(\tilde{x})}, \tag{3.26}$$

and w *and* F *satisfy Eq. 3.22 and Eq. 3.25 for* $\theta^0 = \tilde{\theta}$ *and* $x^0 = \tilde{x}$.

If condition (3.26) is not satisfied, the linear contract can only coordinate either effort level or capacity, and not both. To coordinate the effort level, w *and* F *should satisfy Eq. 3.22 and Eq. 3.25, and to coordinate the capacity,* w *and* F *should satisfy Eq. 3.24 and Eq. 3.25 for* $\theta^0 = \tilde{\theta}$ *and* $x^0 = \tilde{x}$.

Proof. See Appendix. □

3.4.3 Nonlinear Contracts

A set of nonlinear contracts $\{w_n^{NL}(\cdot)\}$ consists of continuous and twice differentiable functions $w_n^{NL}(\cdot)$ in z, which reimburses the health provider $w_n^{NL}(z)$ dollars when z individuals among the n individuals who visited the health provider receive the intervention. For the health provider to implement the social welfare maximizing effort level $\tilde{\theta}$ in the second stage, $\tilde{\theta}$ should be the

unique solution of the recourse problem (3.8). Problem (3.8) has a unique maximum if for any realization of demand, say n, the contract $w_n^{NL}(\cdot)$ is concave in z and satisfies $w_n'^{NL}(\tilde{\theta}n) = c + d'(\tilde{\theta})$. Given these two conditions, the health provider's first-stage problem will be:

$$\max_{x \geq \tilde{\theta}\underline{n}} \Omega_R(x) = -K(x) + \int_{\underline{n}}^{x/\tilde{\theta}} \big(w_n^{NL}(\tilde{\theta}n) - c\tilde{\theta}n - nd(\tilde{\theta})\big) f_N(n)dn + \int_{x/\tilde{\theta}}^{\bar{n}} \big(w_n^{NL}(x) - cx - nd(\frac{x}{n})\big) f_N(n)dn. \tag{3.27}$$

Now the health purchaser should design the contract $\{w_n^{NL}(\cdot)\}$ such that $\tilde{x}$ becomes the unique solution of problem (3.27).

Proposition 15 *When the number of customers is verifiable, the nonlinear contract $\{w_n^{NL}(\cdot)\}$ coordinates the health purchaser-health provider relationship if for any realization of demand, say n:*

1. *$w_n^{NL}(z)$ is strictly concave in z,*
2. *$w_n'^{NL}(\tilde{\theta}n) = c + d'(\tilde{\theta})$,*
3. *$\int_{\tilde{x}/\tilde{\theta}}^{\bar{n}} w_n'^{NL}(\tilde{x}) f_N(n)dn = K'(\tilde{x}) + c(1 - F_N(\frac{\tilde{x}}{\tilde{\theta}})) + \int_{\tilde{x}/\tilde{\theta}}^{\bar{n}} d'(\frac{\tilde{x}}{n}) f_N(n)dn$, and*
4. *the objective function of problem (3.27) is zero at $x = \tilde{x}$.*

Proof. See Appendix. □

Corollary 2 *When the number of customers is verifiable, the nonlinear contract $\{w_n^{NL}(\cdot)\}$ in which $w_n^{NL}(z) = \alpha(n)z^2 + \beta(n)z + \gamma$ coordinates the health purchaser-health provider relationship if for any realization of demand, say n:*

1. *$\alpha(n) = \alpha n$ and $\beta(n) = m - \beta n$, where $\alpha = \frac{K'(\tilde{x}) + \frac{1}{\tilde{\theta}} \int_{\tilde{x}/\tilde{\theta}}^{\bar{n}} d'(\frac{\tilde{x}}{t}) f_N(t)dt}{2 \int_{\tilde{x}/\tilde{\theta}}^{\bar{n}} t(\tilde{x} - \tilde{\theta}t) f_N(t)dt}$, $m = c + d'(\tilde{\theta})$, and $\beta = 2\alpha\tilde{\theta}$, and*
2. *The constant γ is determined such that the objective function of problem (3.27) is zero at $x = \tilde{x}$.*

Proof. Condition 1 is obtained from the conditions 2 and 3 of Proposition 15 for the function $w_n^{NL}(z) = \alpha(n)z^2 + \beta(n)z + \gamma$. Condition 2 is equivalent to the condition 4 of Proposition 15. Also note that $\int_{\tilde{x}/\tilde{\theta}}^{\bar{n}} t(\tilde{x} - \tilde{\theta}t) f_N(t)dt$ is always less than zero, which results in $\alpha < 0$ as required for concavity of $w_n^{NL}(z)$. □

One may also try other functional forms for $w_n^{NL}(\cdot)$ in Proposition 15 ; for instance, exponential form $w_n^{NL}(z) = \gamma + \alpha(n)e^{\beta(n)z}$, or power form $w_n^{NL}(z) = \gamma + \alpha(n)z^{\beta(n)}$. The major

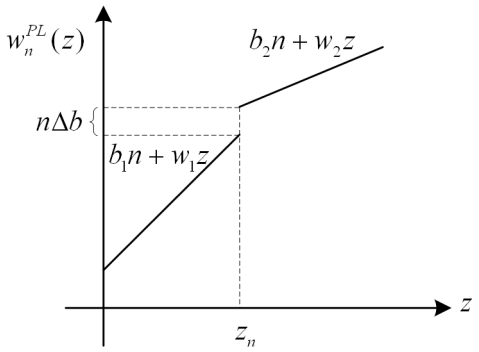


Figure 3.1: Piecewise Linear Contact

drawback of using these functional forms is that one cannot find a closed form solution for the functions $\alpha(n)$ and $\beta(n)$, which results in implementation problems in practice. In the next subsection, we show that piecewise linear contracts can also coordinate the system with less implementation issues than nonlinear contracts.

3.4.4 Piecewise Linear Contracts

Consider a set of piecewise contracts $\{w_n^{PL}(\cdot)\}$ which for any realization of demand, say n, reimburses the health provider according to the following function:

$$w_n^{PL}(z) = \begin{cases} nb_2 + w_2 z, \text{ for } z \geq z_n, \\ nb_1 + w_1 z, \text{ for } z < z_n, \end{cases} \tag{3.28}$$

where b_i and $w_i \geq 0$, $i \in \{0, 1\}$, are constant. Figure 3.1 displays a piecewise contract of form (3.28) for a given n.

For the health provider to implement the social welfare maximizing effort level $\tilde{\theta}$ in the second stage, $\tilde{\theta}$ should be the unique solution of the recourse problem (3.8). It is easy to show (see the proof of Theorem 3) that for the piecewise contract (3.28), problem (3.8) has a unique maximum at $\theta = \tilde{\theta}$ if for any realization of demand, say n, $z_0 = \tilde{\theta} n$, $w_1 > c + d'(\tilde{\theta})$ and $w_2 = c + d'(\tilde{\theta})$. Given these two conditions, the health provider's first-stage problem will be:

$$\begin{aligned} \max_{x \geq \tilde{\theta}\underline{n}} \Omega_R(x) = &- K(x) + \int_{\underline{n}}^{x/\tilde{\theta}} \big(b_1 n + w_1 \tilde{\theta} n + b\Delta n - c\tilde{\theta} n - nd(\tilde{\theta})\big) f_N(n) dn \\ &+ \int_{x/\tilde{\theta}}^{\bar{n}} \big(b_1 n + w_1 x - cx - nd(\frac{x}{n})\big) f_N(n) dn. \end{aligned} \tag{3.29}$$

Now the health purchaser should design the contract $\{w_n^{PL}(\cdot)\}$ such that $\tilde{x}$ becomes the unique solution of problem (3.29).

Theorem 3 *When the number of customers is verifiable, the set of piecewise contracts* $\{w_n^{PL}(\cdot)\}$ *defined according to (3.28) coordinates the health purchaser-health provider relationship if for any realization of demand, say* n*:*

1. $z_n = \tilde{\theta} n$,
2. $w_1 = \frac{K'(\tilde{x}) + \frac{1}{\tilde{\theta}} \int_{\tilde{x}/\tilde{\theta}}^{\bar{n}} d'(\frac{\tilde{x}}{t}) f_N(t) dt - \frac{x}{\tilde{\theta}^2} \Delta b f_N(\frac{\tilde{x}}{\tilde{\theta}})}{1 - F_N(\frac{\tilde{x}}{\tilde{\theta}})} + c$,
3. $w_2 = c + d'(\tilde{\theta})$,
4. $b_1 = \frac{1}{\mathrm{E}[N]} \Big(- (w_1 \tilde{\theta} + \Delta b - c\tilde{\theta} - d(\tilde{\theta})) \int_{\underline{n}}^{\tilde{x}/\tilde{\theta}} t f_N(t) dt$
   $- (w_1 - c) x (1 - F_N(\frac{\tilde{x}}{\tilde{\theta}})) + \int_{\tilde{x}/\tilde{\theta}}^{\bar{n}} t d(\frac{\tilde{x}}{t}) f_N(t) dn + K(\tilde{x}) \Big)$,
5. $\Delta b = \min\{0, \frac{\tilde{\theta}}{\tilde{x} f_N(\frac{\tilde{x}}{\tilde{\theta}})} \Big(K'(\tilde{x}) + \frac{1}{\tilde{\theta}} \int_{\tilde{x}/\tilde{\theta}}^{\bar{n}} d'(\frac{\tilde{x}}{t}) f_N(t) dt - (1 - F_N(\frac{\tilde{x}}{\tilde{\theta}}) d'(\tilde{\theta})) \Big) \}$, *and*
6. $b_2 = b_1 + \Delta b + \tilde{\theta}(w_1 - w_2)$.

Proof. See Appendix. □

An important property of the piecewise linear contract in Theorem 3 is that for a given demand n, the shape of the contract $w_n^{PL}(\cdot)$ does not depend on n (since $w_1, w_2, b_1, b_2, \text{and}, \Delta b$ are all independent of n); in fact, the value of n just shifts the locations of contracts$w_n^{PL}(\cdot)$ to sides and, up and down. This property makes implementation of this contract quite easy, since the health purchaser needs to determine the shape parameters of the contract (i.e., $w_1, w_2, b_1, b_2, \text{and}, \Delta b$) only once, regardless of the value of observed demand.

3.5 Coordinating Contract with Unverifiable Number of Customers for the Medical Intervention

In the previous section, we defined "the number of customers for a medical intervention, N," as the number of individuals who visit the health provider to consume the medical intervention if prescribed by the health provider. We also assumed that any realization of N can be verified by the health purchaser. In practice, however, the number of customers for a medical intervention may not be easily verified by the health purchaser (for instance, refer to Yaesoubi and Roberts (2009) for some examples where the demand verifiability assumption could be undermined). In this section, we relax the verifiability assumption, and characterize the coordinating contracts for health care system where the number of customers for an underlying medical intervention is not verifiable by the health purchaser. Let $w(\cdot)$ denote a contract offered by the health purchaser to the health provider that reimburses the health provider $w(z)$ dollars if the intervention is administered to z individuals.

3.5.1 Linear (Fee-For-Service) Contracts

As in §3.4.2, we define a linear (fee-for-service) contract $w^L(\cdot)$ of the form $w^L(z) = F + wz$, where z is the number of individuals to whom the intervention is administered during the contractual period, $F \geq 0$ is a fixed transfer payment made at the beginning of the contractual period, and $w > c$ is a constant service price.

Proposition 16 *When the number of customers is unverifiable, the linear (fee-for-service) contract $w^L(z) = F + wz$ does not implement the social welfare maximizing effort level $\tilde{\theta}$; it, however, implements the social welfare maximizing capacity level $\tilde{x}$ if and only if:*

$$w = \frac{\int_{\tilde{x}/\tilde{\theta}}^{\bar{n}} d'(\tilde{x}/n) f_N(n) dn + K'(\tilde{x})}{1 - F_N(\tilde{x})} + c, \text{ and} \tag{3.30}$$

$$F = K(\tilde{x}) + \left(w\tilde{\theta} - c\tilde{\theta} - d(\tilde{\theta})\right) \mathrm{E}[N] - \int_{\tilde{x}/\tilde{\theta}}^{\bar{n}} \left((w-c)(\tilde{x} - \tilde{\theta}n) - n(d(\frac{\tilde{x}}{n}) - d(\tilde{\theta}))\right) f_N(n) dn \geq 0. \tag{3.31}$$

Proof. Proof is similar to that of Proposition 14 and hence omitted. □

3.5.2 Nonlinear Contracts

Consider a nonlinear contract $w^{NL}(\cdot)$, which is a continuous and twice-differentiable function. In order to let the recourse problem (3.8) have a unique interior solution, we assume that the contract $w^{NL}(\cdot)$ is concave. Since the health purchaser cannot verify the number of customers, the contract $w^{NL}(\cdot)$ cannot be designed as such to ensure the selection of $\tilde{\theta}$ by the health provider. Therefore, for any realization of N in the second stage, provided that enough capacity is available, the health provider chooses the effort level θ^0 that maximizes his profit. As in §3.4.3, θ^0 will solve:

$$w'^{NL}(n\theta^0) = c + d'(\theta^0). \tag{3.32}$$

By this equation, it is apparent that the effort level θ^0 depends on the number of observed customers, n; to make it explicit, we denote the effort level θ^0 by $\theta^0(n)$.

Similar to §3.4.2, for given x and n, the solution of the recourse problem (3.8) will be as follows:

$$\theta^*(n) = \begin{cases} \theta^0(n), \text{ if } n\theta^0(n) \leq x, \\ x/n, \text{ if } n\theta^0(n) \geq x, \end{cases} \tag{3.33}$$

where $\theta^0(n)$ solves Eq. 3.32.

To find the health provider's first-stage objective function, we should first find the demand n^0, such that for the demand less than n^0, the health provider would have enough capacity

to exert the effort $\theta^0(n)$, and for any demand greater than n^0, the health provider would be bounded by capacity and have to exert the effort level x/n. By (3.33), n^0 solves $n^0\theta^0(n^0) = x$. By substituting $\theta^0(n^0) = x/n^0$ in Eq. (3.32) , we get:

$$n^0(x) = \frac{x}{d'^{-1}(w'(x) - c)} \tag{3.34}$$

Now the health provider's first-stage problem will be:

$$\begin{aligned}\max_{x\geq 0|n^0(x)\geq \underline{n}} \Omega_R(x) = &- K(x) + \int_{\underline{n}}^{n^0(x)} \left(w^{NL}(n\theta^0(n)) - c\theta^0(n)n - nd(\theta^0(n))\right) f_N(n)dn \\ &+ \int_{n^0(x)}^{\bar{n}} \left(w^{NL}(x) - cx - nd(\frac{x}{n})\right) f_N(n)dn. \end{aligned} \tag{3.35}$$

Proposition 17 *When the number of customers is unverifiable, the nonlinear concave contract $w^{NL}(\cdot)$ does not implement the social welfare maximizing effort level $\tilde{\theta}$; it, however, implements the social welfare maximizing capacity level $\tilde{x}$ if*

1. $w'^{NL}(\tilde{x}) = \frac{K'(\tilde{x}) + \int_{n^0(\tilde{x})}^{\bar{n}} d'(\frac{\tilde{x}}{n}) f_N(n)dn}{1 - F_N(n^0(\tilde{x}))} + c$, *where* $n^0(\tilde{x}) = \frac{\tilde{x}}{d'^{-1}(w'(\tilde{x}) - c)}$, *and*
2. *the objective function of problem (3.35) is zero at $x = \tilde{x}$.*

Proof. See Appendix. □

3.5.3 Incentive Feasible Menu of Contracts

The main reason that the contracts studies in §§3.5.1 and 3.5.2 fail to coordinate the system lies in the fact that those contracts do not incent the health provider to reveal the true value of the observed demand. In this section, we characterize a set of contracts that provide enough incentive to the health provider to truthfully reveal the observed demand. For simplicity we assume that the demand can only take two possible values: low and high. The contracts characterized under this setting can be extended to cases where demand can take more possible values. Let's assume that demand can be n_L with probability α_L, and $n_H > n_L$ with probability α_H, where $\alpha_L + \alpha_H = 1$. The health purchaser's problem is to design a menu of contracts $\{w_L^{IF}(\cdot), w_H^{IF}(\cdot)\}$ as to incent the health provider to allocate sufficient capacity, exert optimal effort level, and then reveal the observed demand by selecting the appropriate contract.

The following sequence of events occurs: the health purchaser offers a menu of contracts $\{w_L^{IF}(\cdot), w_H^{IF}(\cdot)\}$; the health provider allocate capacity x; demand will be realized, and the health provider exert effort θ_L if demand is low and effort θ_H if demand is high. The health provider then selects the contract $w_L(\cdot)$ or $w_H(\cdot)$ based on which he would like to be reimbursed.

The health purchaser solves the following optimization problem to find the menu of contracts $\{w_L^{IF}(\cdot), w_H^{IF}(\cdot)\}$ along with the optimal capacity allocation x^* and the effort levels θ_L and θ_H to be exerted when the demand is low and high, respectively.

$$\max_{\substack{0\le\theta_L\le 1, 0\le\theta_H\le 1\\ x\ge 0, w_L^{IF}(\cdot), w_H^{IF}(\cdot)}} \quad \Omega_R = \alpha_L\left(n_L\Pi(\theta_L^0) - w_L^{IF}(\theta_L^0 n_L)\right) + \alpha_H\left(n_H\Pi(\theta_H^0) - w_H^{IF}(\theta_H^0 n_H)\right) \tag{3.36}$$

$$\text{s.t.} \quad \theta_i^0 = \begin{cases} \theta_i, \theta_i n_i \le x, \\ x/n_i, \theta_i n_i \ge x \end{cases}, \text{for} i \in \{L, H\}, \tag{3.37}$$

$$\theta_i = \arg\max_{\theta \le x/n_i}\{w_i^{IF}(\theta n_i) - c\theta n_i - n_i d(\theta)\}, \text{ for } i \in \{L, H\}, \tag{3.38}$$

$$w_i^{IF}(\theta_i^0 n_i) - c\theta_i^0 n_i - n_i d(\theta_i^0) \ge \max_{\theta \le x/n_i}\{w_j^{IF}(\theta n_i) - c\theta n_i - n_i d(\theta)\}, \quad \text{for } j \ne i \in \{L, H\}, \tag{3.39}$$

$$x = \arg\max_{\breve{x}\ge 0} \sum_{i\in\{L,H\}} \alpha_i\left(w_i^{IF}(\theta_i^0 n_i) - c\theta_i^0 n_i - n_i d(\theta_i^0)\right) - K(\breve{x}), \tag{3.40}$$

$$\sum_{i\in\{L,H\}} \alpha_i\left(w_i^{IF}(\theta_i^0 n_i) - c\theta_i^0 n_i - n_i d(\theta_i^0)\right) - K(x) \ge 0. \tag{3.41}$$

Constraints (3.37) and (3.38) determine the optimal effort levels (θ_i^0) in the second-stage for a given capacity x, and the observed demand n_i, for $i \in \{L, H\}$. Constraint (3.40) ensures that the health provider selects the contract $w_i(\cdot)$ if the realized demand is n_i, for $i \in \{L, H\}$. Constraint (3.40) specifies the capacity allocated by the health provider, and constraint (3.41) ensures that the provider accepts the menu of contracts $\{w_L^{IF}(\cdot), w_H^{IF}(\cdot)\}$.

To solve the optimization problem (3.36)-(3.41), we divide the feasible region specified by the constraints (3.37)-(3.41) into three regions. We then find the optimal solution of problem (3.36)-(3.41) over each new region; the global optimum will be the one that results in maximizing the objective function (3.36). To construct the new regions, we add the following constraints to the constraint set (3.37)-(3.41):

Region 1. Add $x \le \min\{n_L\theta_L, n_H\theta_H\}$;

Region 2. Add $\min\{n_L\theta_L, n_H\theta_H\} \le x \le \max\{n_L\theta_L, n_H\theta_H\}$;

Region 3. Add $x \ge \max\{n_L\theta_L, n_H\theta_H\}$.

If the optimal solution falls in Region 1, then apparently the health purchaser does not try to induce any effort level that requires more capacity than x. Therefore Region 1 will be equivalent to a region constructed by adding the constraint $x = n_L\theta_L = n_H\theta_H$ to the constraint set (3.37)-(3.41).

If the optimal solution falls in Region 3, then the health purchaser does not induce a capacity level that may not be entirely utilized under neither n_L nor n_H. Hence, Region 3 is the union of

two sub-regions constructed by adding the constraints $n_L\theta_L \leq n_H\theta_H = x$ or $n_H\theta_H \leq n_L\theta_L = x$ to the constraint set (3.37)-(3.41). These two sub-regions are already included in Region 2, and therefore, Region 3 does not need to be investigated.

If the optimal solution falls in Region 2, then by the same reasoning as in previous paragraphs, this region is equivalent to the union of two sub-regions constructed by adding the constraints $n_L\theta_L \leq n_H\theta_H = x$ or $n_H\theta_H \leq n_L\theta_L = x$ to the constraint set (3.37)-(3.41). Therefore, in order to solve the optimization problem (3.36)-(3.41), we only need to investigate the following scenarios:

Case 1. $x = n_L\theta_L = n_H\theta_H$;

Case 2. $n_L\theta_L \leq n_H\theta_H = x$;

Case 3. $n_H\theta_H \leq n_L\theta_L = x$.

The social welfare resulted in Case 3 is always less than or equal to the social welfare resulted from Case 1. To see this, suppose that at optimum $n_H\theta_H < n_L\theta_L = x$; that is the capacity is totally utilized under low demand but not fully utilized under high demand. This allocation does not maximize the social welfare. To show this, note that under unlimited capacity the effort level at low demand is always equal to the effort level at high demand; therefore, if capacity is fully consumed under low demand, the effort level that maximizes the social welfare under high demand has to consume the entire capacity as well. Therefore, the coordinating contracts obtained in Case 3 will not be of interest to implement; hence, we solve the optimization problem (3.36)-(3.41) for only Case 1 and Case 2.

To find the coordinating contracts under the settings of this subsection, we only consider the contracts of piecewise linear form, which is defined as follows: for $i \in \{L, H\}$,

$$w_i^{IF}(z) = \begin{cases} b_i + w_i z, z \leq z_i^0, \\ b_i^+ + w_i^+ z, z \geq z_i^0. \end{cases} \tag{3.42}$$

As we will see later, the health purchaser can induce any desired level of capacity and effort by determining the parameters of the contract (3.42) appropriately. In what follows, we find the coordinating contract for the cases described above.

Case 1: $x = n_L\theta_L = n_H\theta_H$

When $x = n_L\theta_L = n_H\theta_H$, the objective function (3.36) becomes:

$$\Omega_R = \alpha_L\big(n_L\Pi(\frac{x}{n_L}) - b_L - w_L x\big) + \alpha_H\big(n_H\Pi(\frac{x}{n_H}) - b_H - w_H x\big). \tag{3.43}$$

As we saw in §3.4.4, it is straightforward to show that in order to induce effort levels $\theta_L = x/n_L$ and $\theta_H = x/n_H$, the parameters of contracts (3.42) should satisfy the following

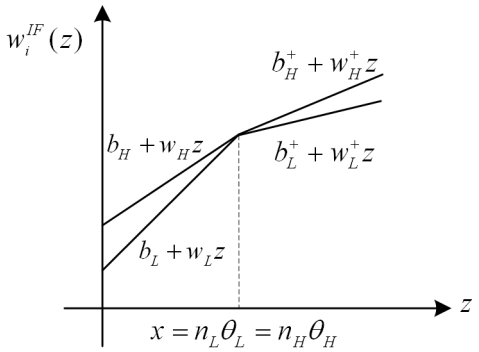


Figure 3.2: Incentive Feasible Menu of Contracts for Case $x = n_L\theta_L = n_H\theta_H$.

conditions:

$$z_i^0 = x = n_L\theta_L = n_H\theta_H \tag{3.44}$$

$$w_L \geq c + d'(x/n_L) \tag{3.45}$$

$$w_H \geq c + d'(x/n_H) \tag{3.46}$$

$$w_L^+ \leq c + d'(x/n_L) \tag{3.47}$$

$$w_H^+ \leq c + d'(x/n_H) \tag{3.48}$$

Figure 3.2 shows two contracts of form (3.42) that satisfy the conditions (3.44)-(3.48). Constraints (3.47) and (3.48) are redundant, since the health provider can never exceed his available capacity x. Therefore, constraints (3.37)-(3.38) are equivalent to the constraints (3.45)-(3.46).

Now suppose that a health provider with demand n_L selects to be reimbursed according to contract $w_H(\cdot)$. He determines his effort level by solving the following problem:

$$\max_{0\leq\theta\leq x/n_L} b_H + w_H\theta n_L - c\theta n_L - n_L d(\theta). \tag{3.49}$$

The problem (3.49) may have an interior or a boundary solution. The first-order derivative of the objective function (3.49) is $n_L\left(w_H - c - d'(\theta)\right)$ which can be either always positive or can become zero at some $\theta \in [0, x/n_L]$.

If $w_H \geq c + d'(x/n_L)$, then $n_L\left(w_H - c - d'(\theta)\right) \geq 0$ for any $\theta \in [0, x/n_L]$, and hence the health provider select the effort level x/n_L. Under this situation, for $i = L$, constraint (3.40) will be equivalent to:

$$b_L + w_L x - cx - n_L d(x/n_L) \geq b_H + w_H x - cx - n_L d(x/n_L).$$

Therefore, if at optimality $w_H \geq c + d'(x/n_L)$, the optimization problem (3.36)-(3.41)

becomes:

$$\max_{x,w_L,w_H\in\Re^+} \Omega_R = \alpha_L\big(n_L\Pi(\frac{x}{n_L}) - b_L - w_L x\big) + \alpha_H\big(n_H\Pi(\frac{x}{n_H}) - b_H - w_H x\big) \tag{OP1}$$

$$\text{s.t.} \quad x \le n_H \tag{3.50}$$

$$w_L \ge c + d'(x/n_L) \tag{3.51}$$

$$w_H \ge c + d'(x/n_L) \tag{3.52}$$

$$b_L + w_L x - cx - n_L d(x/n_L) \ge b_H + w_H x - cx - n_L d(x/n_L) \tag{3.53}$$

$$b_H + w_H x - cx - n_H d(x/n_H) \ge b_L + w_L x - cx - n_H d(x/n_H) \tag{3.54}$$

$$\alpha_L\left(w_L - c - d'(x/n_L)\right) + \alpha_H\left(w_H - c - d'(x/n_H)\right) - K'(x) = 0 \tag{3.55}$$

$$\sum_{i\in\{L,H\}} \alpha_i\left(b_i + w_i x - cx - n_i d(x/n_i)\right) - K(x) \ge 0. \tag{3.56}$$

In optimization problem OP1, constraints (3.51)-(3.52) correspond to constraints (3.37)-(3.38), constraints (3.53)-(3.54) correspond to constraint set (3.40), constraint (3.55) represents the optimality condition of selecting capacity x (i.e., constraint (3.40)), and constraint (3.56) corresponds to the participation constraint (3.41).

Now if at optimality $w_H \le c + d'(x/n_L)$, then along with constraint (3.46), $d'(x/n_H) - d'(\theta) \le w_H - c - d'(\theta) \le d'(x/n_L) - d'(\theta)$. It implies that for $\theta = 0$,
$n_L\left(w_H - c - d'(\theta)\right) \ge 0$ and for $\theta = x/n_L$, $n_L\left(w_H - c - d'(\theta)\right) \le 0$; hence the derivative of the objective function (3.49) reaches zero at one and only one point over the interval $[0, x/n_L]$, and therefore the optimization problem (3.49) is maximized at point $\theta^* = d'^{-1}(w_H - c)$. Under this situation, for $i = L$, constraint (3.40) will be equivalent to:

$$b_L + w_L x - cx - n_L d(x/n_L) \ge b_H + (w_H - c)n_L d'^{-1}(w_H - c) - n_L d\left(d'^{-1}(w_H - c)\right)$$

Now suppose that a health provider with demand n_H selects to be reimbursed according to contract $w_L(\cdot)$. He determines his effort level by solving the following problem:

$$\max_{0\le\theta\le x/n_H} b_L + w_L\theta n_H - c\theta n_H - n_H d(\theta). \tag{3.57}$$

The first-order derivative of the objective function (3.57) is $n_H\left(w_L - c - d'(\theta)\right)$ which is greater than or equal to $n_H\left(d'(x/n_L) - d'(\theta)\right)$ by (3.45); and hence always positive for any $\theta \in [0, x/n_H]$. Therefore, for $i = H$, constraint (3.40) will be equivalent to:

$$b_H + w_H x - cx - n_H d(x/n_H) \ge b_L + w_L x - cx - n_H d(x/n_H).$$

Given that $x = n_L\theta_L = n_H\theta_H$, constraint (3.40) becomes:

$$x = \arg\max_{\breve{x}\geq 0} \sum_{i\in\{L,H\}} \alpha_i \left(b_i + w_i\breve{x} - c\breve{x} - n_i d(\breve{x}/n_i)\right) - K(\breve{x}),$$

with necessary and sufficient optimality condition:

$$\alpha_L \left(w_L - c - d'(x/n_L)\right) + \alpha_H \left(w_H - c - d'(x/n_H)\right) - K'(x) = 0.$$

Therefore, if at optimality $w_H \leq c + d'(x/n_L)$, the optimization problem (3.36)-(3.41) becomes equivalent to:

$$\max_{x,w_L,w_H\in\Re^+} \Omega_R = \alpha_L\big(n_L\Pi(\frac{x}{n_L}) - b_L - w_L x\big) + \alpha_H\big(n_H\Pi(\frac{x}{n_H}) - b_H - w_H x\big) \tag{OP2}$$

$$\text{s.t.} \quad x \leq n_H \tag{3.58}$$

$$w_L \geq c + d'(x/n_L) \tag{3.59}$$

$$w_H \geq c + d'(x/n_H) \tag{3.60}$$

$$w_H \leq c + d'(x/n_L) \tag{3.61}$$

$$b_L + w_L x - cx - n_L d(x/n_L) \geq b_H + (w_H - c)n_L d'^{-1}(w_H - c) - n_L d\left(d'^{-1}(w_H - c)\right) \tag{3.62}$$

$$b_H + w_H x - cx - n_H d(x/n_H) \geq b_L + w_L x - cx - n_H d(x/n_H) \tag{3.63}$$

$$\alpha_L \left(w_L - c - d'(x/n_L)\right) + \alpha_H \left(w_H - c - d'(x/n_H)\right) - K'(x) = 0 \tag{3.64}$$

$$\sum_{i\in\{L,H\}} \alpha_i \left(b_i + w_i x - cx - n_i d(x/n_i)\right) - K(x) \geq 0. \tag{3.65}$$

In optimization problem OP2, constraints (3.59)-(3.61) correspond to constraints (3.37)-(3.38), constraints (3.62)-(3.63) correspond to constraint set (3.40), constraint (3.64) represents the optimality condition of selecting capacity x (i.e., constraint (3.40)), and constraint (3.65) corresponds to the participation constraint (3.41).

Thus for Case 1, the health purchaser solves optimization problems OP1 and OP2, and selects the solution that maximizes her objective function.

Case 2: $n_L\theta_L \leq n_H\theta_H = x$

When $n_L\theta_L \leq n_H\theta_H = x$, the objective function (3.36) becomes:

$$\Omega_R = \alpha_L\big(n_L\Pi(\theta_L) - b_L - w_L\theta_L n_L\big) + \alpha_H\big(n_H\Pi(\frac{x}{n_H}) - b_H - w_H x\big). \tag{3.66}$$

As for Case 1, it is straightforward to show that in order to induce effort levels θ_L and

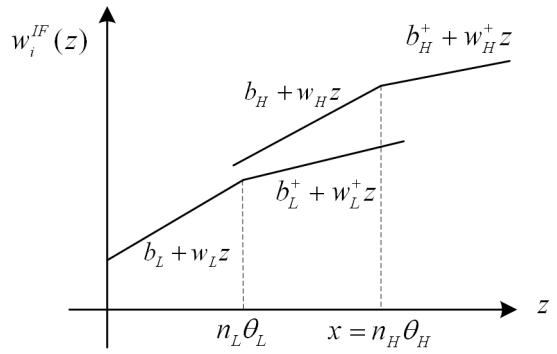


Figure 3.3: Incentive Feasible Menu of Contracts for Case $n_L\theta_L \leq n_H\theta_H = x$

$\theta_H = x/n_H$, the parameters of contracts (3.42) should satisfy the following conditions:

$$z_i^0 = n_i\theta_i \text{ for } i \in \{L, H\} \tag{3.67}$$

$$w_L \geq c + d'(\theta_L) \tag{3.68}$$

$$w_H \geq c + d'(x/n_H) \tag{3.69}$$

$$w_L^+ \leq c + d'(\theta_L) \tag{3.70}$$

$$w_H^+ \leq c + d'(x/n_H) \tag{3.71}$$

Figure 3.3 shows two contracts of form (3.42) that satisfy the conditions (3.67)-(3.71). Constraints (3.71) is redundant, since the health provider can never exceed his available capacity x. Therefore, constraints (3.37)-(3.38) are equivalent to the constraints (3.67)-(3.71).

Now suppose that a health provider with demand n_L selects to be reimbursed according to contract $w_H(\cdot)$. He determines his effort level by solving the problem (3.49). This problem may again have an interior or a boundary solution, as described in Case 1. Therefore, if $w_H \geq c + d'(x/n_L)$, constraint (3.40) will be equivalent to:

$$b_L + w_L\theta_L n_L - c\theta_L n_L - n_L d(\theta_L) \geq b_H + w_H x - cx - n_L d(x/n_L).$$

And if $w_H \leq c + d'(x/n_L)$ it will be equivalent to:

$$b_L + w_L\theta_L n_L - c\theta_L n_L - n_L d(\theta_L) \geq b_H + (w_H - c)n_L d'^{-1}(w_H - c) - n_L d\left(d'^{-1}(w_H - c)\right).$$

Now suppose that a health provider with demand n_H selects to be reimbursed according to contract $w_L(\cdot)$. He determines his effort level by solving the following problem:

$$\max_{0 \leq \theta \leq x/n_H} w_L(\theta n_H) - c\theta n_H - n_H d(\theta).$$

It is easy to see that by conditions (3.68) and setting $w_L^+ \leq c + d'(\theta_L n_L/n_H)$, the health provider with demand n_H selects the effort level θ_L. Therefore, for $i = H$, constraint (3.40) will be equivalent to:

$$b_H + w_H x - cx - n_H d(x/n_H) \geq b_L + w_L \theta_L n_L - c\theta_L n_L - n_H d(\theta_L n_L/n_H).$$

Given that $n_L\theta_L \leq n_H\theta_H = x$, constraint (3.40) becomes equivalent to:

$$\alpha_H \left(w_H - c - d'(x/n_H)\right) - K'(x) = 0 \text{ for } x \geq \theta_L n_L.$$

Therefore, for Case 2, the health purchaser solves the following two optimization problems, and selects the solution that maximizes her objective function. If $w_H \geq c + d'(x/n_L)$, then the optimization problem (3.36)-(3.41) becomes equivalent to:

$$\max_{\substack{0\leq\theta_L\leq 1,\\ x,w_L,w_L^+,w_H\in\Re^+}} \quad \Omega_R = \alpha_L\big(n_L\Pi(\theta_L) - b_L - w_L\theta_L n_L\big) + \alpha_H\big(n_H\Pi(\frac{x}{n_H}) - b_H - w_H x\big) \qquad \text{(OP3)}$$

$$\text{s.t.} \quad x \geq \theta_L n_L \tag{3.72}$$

$$x \leq n_H \tag{3.73}$$

$$w_L \geq c + d'(\theta_L) \tag{3.74}$$

$$w_L^+ \leq c + d'(\theta_L n_L/n_H) \tag{3.75}$$

$$w_H \geq c + d'(x/n_L) \tag{3.76}$$

$$b_L + w_L\theta_L n_L - c\theta_L n_L - n_L d(\theta_L) \geq b_H + w_H x - cx - n_L d(x/n_L) \tag{3.77}$$

$$b_H + w_H x - cx - n_H d(x/n_H) \geq b_L + w_L\theta_L n_L - c\theta_L n_L - n_H d(\theta_L n_L/n_H) \tag{3.78}$$

$$\alpha_H \left(w_H - c - d'(x/n_H)\right) - K'(x) = 0 \tag{3.79}$$

$$\alpha_L \left(b_L + w_L\theta_L n_L - c\theta_L n_L - n_L d(\theta_L)\right) + \alpha_H \left(b_H + w_H x - cx - n_H d(x/n_H)\right) - K(x) \geq 0 \tag{3.80}$$

In optimization problem OP3, constraints (3.74)-(3.76) correspond to constraints (3.37)-(3.38), constraints (3.77)-(3.78) correspond to constraint set (3.40), constraint (3.79) represents the optimality condition of selecting capacity x (i.e., constraint (3.40)), and constraint (3.80) corresponds to the participation constraint (3.41).

If at optimum, $w_H \leq c + d'(x/n_L)$, then the optimization problem (3.36)-(3.41) becomes equivalent to:

$$\max_{\substack{0\le\theta_L\le 1,\\ x,w_L,w_L^+,w_H\in\Re^+}} \quad \Omega_R = \alpha_L\big(n_L\Pi(\theta_L) - b_L - w_L\theta_L n_L\big) + \alpha_H\big(n_H\Pi(\frac{x}{n_H}) - b_H - w_H x\big) \qquad \text{(OP4)}$$

$$\text{s.t.} \quad x \ge \theta_L n_L \tag{3.81}$$

$$x \le n_H \tag{3.82}$$

$$w_L \ge c + d'(x/n_L) \tag{3.83}$$

$$w_L^+ \le c + d'(\theta_L n_L/n_H) \tag{3.84}$$

$$w_H \ge c + d'(x/n_H) \tag{3.85}$$

$$w_H \le c + d'(x/n_L) \tag{3.86}$$

$$b_L + w_L\theta_L n_L - c\theta_L n_L - n_L d(\theta_L) \ge b_H + (w_H - c)n_L d'^{-1}(w_H - c) - n_L d\left(d'^{-1}(w_H - c)\right) \tag{3.87}$$

$$b_H + w_H x - cx - n_H d(x/n_H) \ge b_L + w_L\theta_L n_L - c\theta_L n_L - n_H d(\theta_L n_L/n_H) \tag{3.88}$$

$$\alpha_H\left(w_H - c - d'(x/n_H)\right) - K'(x) = 0 \tag{3.89}$$

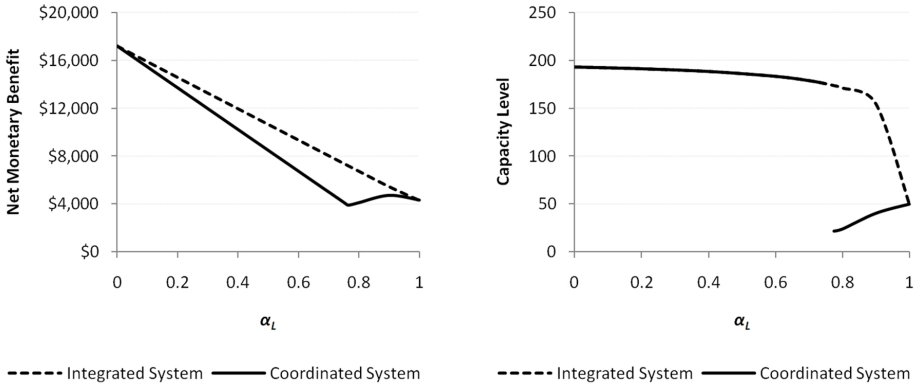
$$\alpha_L\left(b_L + w_L\theta_L n_L - c\theta_L n_L - n_L d(\theta_L)\right) + \alpha_H\left(b_H + w_H x - cx - n_H d(x/n_H)\right) - K(x) \ge 0 \tag{3.90}$$

In optimization problem OP4, constraints (3.83)-(3.86) correspond to constraints (3.37)-(3.38), constraints (3.87)-(3.88) correspond to constraint set (3.40), constraint (3.89) represents the optimality condition of selecting capacity x (i.e., constraint (3.40)), and constraint (3.90) corresponds to the participation constraint (3.41).

Numerical Analysis

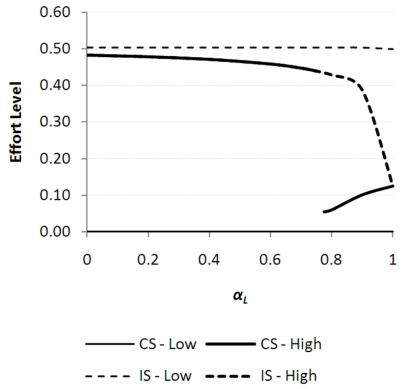
The optimization problems OP1-OP4 are well-behaved for a wide range of parameter values; that is, their global optimums can be reached through a reasonable number of iterations. We used Lingo (Schrage 2008) to obtain the numerical results of this subsection. For all the optimization problems solved to generate the results presented here, Lingo was able to identify the *global* solution; and in 95% of cases the global solution was found in less than 1 second.

To determine the model parameters, we normalized the variable cost c to \$1; and then specified the functions $\Pi(\cdot)$, $d(\cdot)$ and $K(\cdot)$ as such to yield reasonable values for the social welfare maximizing effort level $\tilde{\theta}$ and capacity level $\tilde{x}$ (for instance $0.25 \le \tilde{\theta} \le 0.75$, $1.25n_L \le \tilde{x} \le 0.75n_H$). We then investigate the capability of the incentive feasible coordinating contract in maximizing the social welfare for different demand distributions $(n_L, \alpha_L; n_H, \alpha_H)$. We conduct the analysis for the following two systems:



(a) Net Monetary Benefit vs. a_L

(b) Capacity (x) vs. a_L



(c) Effort Level (θ) vs. a_L (CS: Coordinated system, IS: Integrated system, Low: Low demand, High: High demand

Figure 3.4: Effect of α_L on system coordination

1. **Integrated system**, in which the health purchaser and the health provider are in perfect cooperation as such the health purchaser has control over the health provider's capacity and effort level decisions.

2. **Coordinated system**, in which the health purchaser and the health provider are disjoint organizations but the health purchaser offers a set of incentive feasible coordinating contracts to coordinate their interaction.

Figure 3.4 displays the health purchaser's net monetary benefit (NMB), allocated capacity level, and effort level for the systems described above, assuming that $n_L = 100$ and $n_H = 400$. Remember that for the integrated system, the health purchaser's net monetary benefit is the social welfare function (3.1), and for the coordinated system, the health purchaser's net monetary benefit is the objective function (3.36).

Figure 3.4(a) shows the health purchaser's NMB for both systems versus α_L. Apparently, the health purchaser's NMB in both systems is maximized when $\alpha_L = 0$ (or $\alpha_H = 1$), since the health purchaser is faced with no uncertainty and also a larger number of people are using the preventive intervention (demand is n_H). The difference between the net monetary benefit functions are in fact the *information rent* that the health purchaser has to pay to the health provider to select a contract that *reveal* his true demand. This difference is maximum at $\alpha_L = 0.7625$; this is the point in which the health purchaser reaches its lowest capability in distinguishing the type of the health providers (low demand versus high demand). After this point, the health purchaser again gains his capability in coordinating the system, which results in higher NMB.

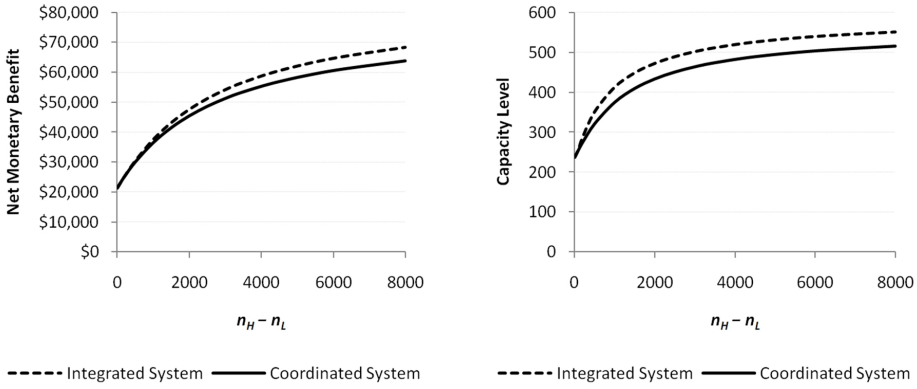
Figure 3.4(b) and Figure 3.4(c) show the optimal effort level and capacity level for both the integrated and coordinated systems. For $\alpha_L \in [0, 0.7625]$, the health purchaser induces a capacity level which is exactly equal to the social maximizing capacity level $\tilde{x}$. However, to distinguish between health providers, the health purchaser induces almost zero effort level for the low-demand health provider. At $\alpha_L = 0.7625$, reaching her lowest capability in inferring the true demand, the health purchaser selects to induce a very low capacity level. After this point, the health purchaser begins inducing higher capacity, and higher effort level for the high-demand health provider; however, the effort level for the low-demand health provider still remains very close to zero.

In the second part of the numerical analysis, we investigate the effect of the distance between n_L and n_H on the system coordination. Figure 3.5 displays the health purchaser's net monetary benefit (NMB), allocated capacity level, and effort level for the systems described above, assuming that $\alpha_L = \alpha_H = 0.5$ and $n_L = 500$.

Figure 3.5(a) shows the health purchaser's NMB versus $n_H - n_L$. As $n_H - n_L$ approaches zero, the health purchaser's NMB in a coordinated system gets closer to her NMB in an integrated system, since under this situation, she is faced with minimal uncertainty in the demand. However, as $n_H - n_L$ increases, the health purchaser's NMB in a coordinated systems distances more from her NMB in an integrated systems. This is due to the fact that in order to distinguish between the health providers with low and high demand, the health purchaser selects to induce effort and capacity levels that deviate from the welfare maximizing allocation (as shown in Figure 3.5(b) and Figure 3.5(c)), which consequently results in an decrease in the health purchaser's NMB.

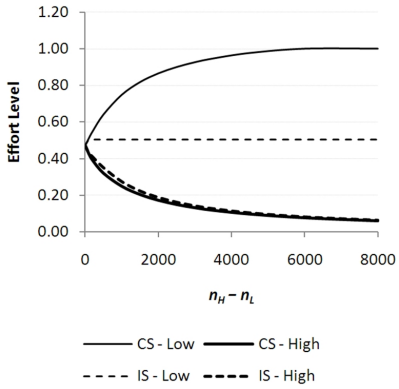
3.6 Conclusion and Future Research

In this paper, we studied the contracts that coordinate the health purchaser-health provider relationship in a preventive health care delivery system where the health provider is limited



(a) Net Monetary Benefit vs. $n_H - n_L$

(b) Capacity (x) vs. $n_H - n_L$



(c) Effort Level (θ) vs. a_L (CS: Coordinated system, IS: Integrated system, Low: Low demand, High: High demand

Figure 3.5: Effect of $n_H - n_L$ on system coordination

by capacity in providing the medical intervention. Such contracts allow the health purchaser and the health provider to optimize their objective functions while maximizing the population welfare. The proposed principal-agent model considers both the problem of moral hazard (hidden action) and asymmetric information (hidden information). In this model, the health provider's decision about the capacity level to allocate and the rank of patients to whom the preventive intervention should be administered is not observable (hidden action). The health purchaser may also be unable to observe the number of customers for the medical intervention (hidden information).

When the number of customers for a medical intervention is verifiable by the health purchaser, we show that gate-keeping, fee-for-service, and nonlinear contracts do not necessarily coordinate health purchaser-health provider relationship. We demonstrated that a set of piecewise linear contacts can coordinate the system under this assumption. When the number of

customers is not verifiable, we showed that the fee-for-service and nonlinear contracts do not coordinate the health purchaser-health provider relationship. Under this setting, the coordinating contract can be a menu of incentive-feasible piecewise linear contracts.

Our model assumes that there is only one medical intervention available for the underlying disease. An immediate extension to our model will be finding the coordinating contracts for cases where more than one medical intervention is available for the disease. In the multiple-intervention paradigm, the coordinating contracts lead the health provider to prescribe the intervention for each patient that results in the best outcome. Also, when demand is unverifiable, we characterized the coordinating contracts when the demand can only take two values. An interesting topic for future research is to find the coordinating contracts when demand takes n values.

Appendices

A.1. Proof of Proposition 12

First we need to show that a positive social welfare maximizing capacity level ($\tilde{x}$) cannot be less than $\tilde{\theta}\underline{n}$ (note that the social welfare maximizing effort level $\tilde{\theta}$ satisfies $\Pi'(\tilde{\theta}) = c + d'(\tilde{\theta})$) . To show this, suppose the contrary; that is $0 < \tilde{x} < \tilde{\theta}\underline{n}$. Hence the optimization problem (3.2) becomes:

$$\max_{0\leq x<\tilde{\theta}\underline{n}} \Omega(x) = -K(x) + \int_{\underline{n}}^{\bar{n}} \left(n\Pi(\frac{x}{n}) - cx - nd(\frac{x}{n})\right) f_N(n)dn. \tag{3.91}$$

The first derivative of the objective function in (3.91) is:

$$\frac{\partial}{\partial x}\Omega(x) = -K'(x) + \int_{\underline{n}}^{\bar{n}} \left(\Pi'(\frac{x}{n}) - c - d'(\frac{x}{n})\right) f_N(n)dn. \tag{3.92}$$

The social welfare maximizing effort level $\tilde{\theta}$ satisfies $\Pi'(\tilde{\theta}) = c + d'(\tilde{\theta})$; substituting $c = \Pi'(\tilde{\theta}) - d'(\tilde{\theta})$ in (3.92) results in:

$$\frac{\partial}{\partial x}\Omega(x) = -K'(x) + \int_{\underline{n}}^{\bar{n}} \left(\Pi'(\frac{x}{n}) - \Pi'(\tilde{\theta}) - (d'(\frac{x}{n}) - d'(\tilde{\theta}))\right) f_N(n)dn. \tag{3.93}$$

Since we assumed that $\tilde{x} < \theta^0\underline{n}$; therefore $\tilde{x} < \theta^0 n$ for any $n \in [\underline{n}, \bar{n}]$. Consequently, since $\Pi(\cdot)$ is strictly concave, and $d(\cdot)$ and $K(\cdot)$ are strictly convex, the first derivative (3.93) is always less than zero for any $x < \tilde{\theta}\underline{n}$. Therefore the objective function (3.91) is maximized if $x = 0$; this contradicts the original assumption; thus $\tilde{x} \geq \tilde{\theta}\underline{n}$.

Condition (3.6) is the optimality condition of recourse function (3.3). The first-order optimality condition for problem (3.4) is:

$$\frac{\partial}{\partial x}\Omega(x) = -K'(x) + \int_{x/\theta^0}^{\bar{n}} \left(\Pi'(\frac{x}{n}) - d'(\frac{x}{n})\right) f_N(n)dn - c\left(1 - F_N(\frac{x}{\theta^0})\right) = 0.$$

By concavity of $\Pi(\cdot)$, and the fact that θ^0 solves $\Pi'(\theta^0) = c + d'(\theta^0)$, we have:

$$\frac{\partial^2}{\partial x^2}\Omega(x) = -K''(x) + \int_{x/\theta^0}^{\bar{n}} \frac{1}{n}\left(\Pi''(\frac{x}{n}) - d''(\frac{x}{n})\right) f_N(n)dn < 0$$

Therefore, the first-order condition is also the sufficient condition for optimality. The result is then immediate. □

A.2. Proof of Proposition 14

To prove, first note that condition (3.24) is the sufficient condition of optimality for the health provider's first-stage problem (3.23); to see this, since $\Pi'(\theta^0) = c + d'(\theta^0)$, the second derivative

of objective function (3.23) satisfies:

$$\frac{\partial^2}{\partial x^2}\Omega_R(x) = -K''(x) - \int_{x/\theta^0}^{\bar{n}} \frac{1}{n} d''(\frac{x}{n}) f_N(n) dn < 0$$

To coordinate the system, $(\tilde{w}, \tilde{F})$ should satisfy the first-order optimality conditions of both the first and the second stages (i.e., conditions (3.22) and (3.24), respectively). This requirement is satisfied if and only if Eq. (3.26) holds. Moreover, $(\tilde{w}, \tilde{F})$ should satisfy the binding participation constraint (3.25); this gives the proof of the first part.

For the second part, where condition (3.26) is not satisfied, the contract can only coordinate either effort level or capacity, and not both since no $(\tilde{w}, \tilde{F})$ can satisfy both the conditions (3.22) and (3.24) simultaneously. □

A.3. Proof of Proposition 15

Conditions 1 and 2 are required to make the social maximizing effort level $\tilde{\theta}$ the unique solution of the recourse problem (3.8). Condition 3 is obtained from simplifying the first-order optimality condition of problem (3.27) which is:

$$-K'(\tilde{x}) + \int_{\tilde{x}/\tilde{\theta}}^{\bar{n}} (w_n'^{NL}(\tilde{x}) - c - d'(\frac{\tilde{x}}{n})) f_N(n) dn = 0. \tag{3.94}$$

To show that the first-order condition (3.94) is also sufficient condition for optimality, we calculate the second-order derivative of the objective function (3.27) at $\tilde{x}$:

$$\frac{\partial^2}{\partial x^2}\Omega_R(\tilde{x}) = -K''(\tilde{x}) - \frac{1}{\tilde{\theta}}(w_{\tilde{x}/\tilde{\theta}}'^{NL}(\tilde{x}) - c - d'(\frac{\tilde{x}}{n})) f_N(\frac{\tilde{x}}{\tilde{\theta}}) + \int_{\tilde{x}/\tilde{\theta}}^{\bar{n}} (w_n''^{NL}(\tilde{x}) - d''(\frac{\tilde{x}}{n})) f_N(n) dn.$$

Since setting $n = \tilde{x}/\tilde{\theta}$ in condition 2 results in $w_{\tilde{x}/\tilde{\theta}}'^{NL}(\tilde{x}) - c - d'(\tilde{\theta}) = 0$, we will have $\frac{\partial^2}{\partial x^2}\Omega_R(\tilde{x}) < 0$. □

A.4. Proof of Theorem 3

First we need to show that for any realization of demand, the piecewise linear contract implement $\tilde{\theta}$ in the second stage if $w_1 > c + d'(\tilde{\theta})$ and $w_2 = c + d'(\tilde{\theta})$. For any given capacity level x and the realization of demand n, the health provider solves the following optimization problem:

$$\max \quad w_n^{PL}(\theta n) - c\theta n - nd(\theta) \tag{3.95}$$

$$\text{s.t.} \quad \theta n \leq x. \tag{3.96}$$

For any given n, two cases may occur: (1) $\tilde{\theta} n \leq x$ or (2) $\tilde{\theta} n > x$.

Case 1: $\tilde{\theta}n \leq x$ (the health provider has enough capacity to exert the effort level $\tilde{\theta}$): In this case, if the health provider selects an effort level $\theta < \tilde{\theta}$, the derivative of the objective function in (3.95) at point $\theta < \tilde{\theta}$ will be equal to $nw_1 - cn - nd'(\theta) > n(d'(\tilde{\theta}) - d'(\theta)) > 0$.

Therefore, the health provider can increase his gain by increasing his effort level to $\tilde{\theta}$. Now if the health provider selects an effort level the derivative of the objective function in (3.95) at point $\theta > \tilde{\theta}$ will be equal to $nw_2 - cn - nd'(\theta) = n(d'(\tilde{\theta}) - d'(\theta)) < 0$. Hence, the health provider will not be willing to exert an effort greater than $\tilde{\theta}$.

Case 2: $\tilde{\theta}n > x$ (the health provider does not have enough capacity to exert the effort level $\tilde{\theta}$): In this case, the welfare maximizing effort is the one that utilize the capacity entirely, i.e., $\theta = x/n$. If the health exert an effort level $\theta < x/n$, the derivative of the objective function in (3.95) at this point will be equal to $nw_1 - cn - nd'(\theta) > n(d'(\tilde{\theta}) - d'(\theta)) > 0$. And hence the health provider select can maximize his gain by increasing his effort level to $\theta = x/n$.

Now, we need to show that the piecewise linear contract can also coordinate the capacity in the first stage. From the objective function (3.29) we have:

$$\Omega'_R(x) = -K'(x) + \frac{x\Delta b}{\tilde{\theta}^2} f_N(\frac{x}{\tilde{\theta}}) + (w_1 - c)(1 - F_N(\frac{x}{\tilde{\theta}})) - \int_{x/\tilde{\theta}}^{\bar{n}} d'(\frac{x}{n}) f_N(n) dn.$$

For the health provider to select the desired capacity level $\tilde{x}$, $\tilde{x}$ should satisfy the first-order optimality condition $\Omega'_R(\tilde{x}) = 0$, which results in condition 1 of Theorem 1. Also we have:

$$\Omega''_R(x) = -K''(x) + \frac{\Delta b}{\tilde{\theta}^2} f_N(\frac{x}{\tilde{\theta}}) + \frac{x\Delta b}{\tilde{\theta}^3} f'_N(\frac{x}{\tilde{\theta}}) - (w_1 - c - d'(\tilde{\theta}))\frac{1}{\tilde{\theta}} f_N(\frac{x}{\tilde{\theta}}) - \int_{x/\tilde{\theta}}^{\bar{n}} \frac{1}{n} d''(\frac{x}{n}) f_N(n) dn.$$

By conditions 1 and 4 in Theorem 1, $\Omega''_R(x) < 0$; and therefore the first-order condition will be also the sufficient condition for optimality. Condition 4 ensures that the second-stage optimality requirement $w_1 > c + d'(\tilde{\theta})$ is satisfied; condition 3 is derived from the participation constraint (3.13); and condition 5 is by the definition of the contract at point $z = \tilde{\theta}n$ (see Figure 3.1), which requires $nb_1 + \tilde{\theta}nw_1 + n\Delta b = nb_2 + \tilde{\theta}nw_2$. □

A.5. Proof of Proposition 17

The first-order optimality condition of problem (3.35) is:

$$\begin{aligned}\Omega_R'(x) = &- K'(x) \\ &+ \frac{d}{dx}n^0(x) \cdot \Big(w^{NL}\left(n^0(x) \cdot \theta^0\left(n^0(x)\right)\right)\Big) f_N(n^0(x)) \\ &- \frac{d}{dx}n^0(x) \cdot \Big(w^{NL}\left(x\right) - cx - n^0(x) \cdot d\big(\frac{x}{n^0(x)}\big)\Big) f_N(n^0(x)) \\ &- \frac{d}{dx}n^0(x) \cdot \Big(c \cdot \theta^0\left(n^0(x)\right) \cdot n^0(x) - n^0(x) \cdot d\left(\theta^0\left(n^0(x)\right)\right)\Big) f_N(n^0(x)) \\ &\int_{n^0(x)}^{\bar{n}} \Big(w'^{NL}(x) - c - d'(\frac{x}{n})\Big) f_N(n) dn = 0. \end{aligned} \tag{3.97}$$

Note that we have by the definition of $n^0(x)$ and $\theta^0(n)$ we have $n^0(x) \cdot \theta^0\left(n^0(x)\right) = x$. Therefore the first-order condition (3.97) reduces to:

$$\Omega_R'(x) = -K'(x) + \int_{n^0(x)}^{\bar{n}} \Big(w'^{NL}(x) - c - d'(\frac{x}{n})\Big) f_N(n) dn = 0. \tag{3.98}$$

Consequently, we have:

$$\Omega_R''(x) = -K''(x) + \int_{n^0(x)}^{\bar{n}} \Big(w''^{NL}(x) - d''(\frac{x}{n})\Big) f_N(n) dn,$$

which is always negative. Therefore, the first-order optimality condition (3.98) is the necessary and sufficient condition of optimality.

REFERENCES

American Cancer Society. 2007. *Cancer Prevention & Early Detection Facts & Figures*. American Cancer Society, Atlanta, GA.

Baron, Richard J., Christine K. Cassel. 2008. 21st-century primary care: New physician roles need new payment models. *Journal of American Medical Association* **299**(13) 1595–1597.

Bayoumi, Ahmed M., Donald A. Redelmeier. 2000. Decision analysis with cumulative prospect theory. *Medical Decision Making* **20** 404–412.

Bernstein, Fernando, Gregory A. DeCroix. 2004. Decentralized pricing and capacity decisions in a multitier system with modular assembly. *Management Science* **50**(9) 1293–1308.

Blomqvist, Ake. 1997. Optimal non-linear health insurance. *Journal of Health Economics* **16** 303–321.

Boadway, Robin, Maurice Marchand, Motohiro Sato. 2004. An optimal contract approach to hospital financing. *Journal of Health Economics* **23** 85110.

Byrne, Margaret M., Kimberly O'Malley, Maria E. Suarez-Almazor. 2005. Willingness to pay per quality-adjusted life year in a study of knee osteoarthritis. *Medical Decision Making* **25** 655–666.

Cachon, Grard P. 2003. *Handbooks in Operations Research and Management Science: Supply Chain Management*, chap. 6: Supply Chain Coordination with Contracts. North-Holland, 229–340.

Cachon, Grard P., Martin A. Lariviere. 2005. Supply chain coordination with revenue-sharing contracts: Strengths and limitations. *Management Science* **51**(1) 3044.

Centers for Disease Control and Prevention. 2008. *Colorectal (Colon) Cancer*. Centers for Disease Control and Prevention. Available via http://www.cdc.gov/cancer/colorectal/basic_info/index.htm [accessed January 1, 2010].

Centers for Disease Control and Prevention. 2010. *Colorectal (Colon) Cancer: Insurance and Medicare*. Centers for Disease Control and Prevention. Available via http://www.cdc.gov/cancer/colorectal/basic_info/screening/insurance.htm [accessed January 1, 2010].

Chakravarty, Amiya K., Jun Zhang. 2007. Collaboration in contingent capacities with information asymmetry. *Naval Research Logistics* **54** 421–432.

Chalkley, Martin, James M. Malcomson. 1998. Contracting for health services when patient demand does not reflect quality. *Journal of Health Economics* **17** 1–19.

Chernew, Michael E., William E. Encinosa, Richard A. Hirth. 2000. Optimal health insurance: the case of observable, severe illness. *Journal of Health Economics* **19** 585609.

Cookson, Richard. 2003. Willingness to pay methods in health care: a sceptical view. *Health Economics* **13** 891–894.

Diener, Alan, Bernie Obrien, Amiram Gafni. 1998. Health care contingent valuation studies: a review and classification of the literature. *Health Economics* **7** 313–326.

Ellis, Randall P., Willard G. Manning. 2007. Optimal health insurance for prevention and treatment. *Journal of Health Economics* **26** 11281150.

Frew, E., J.L Wolstenholme, D. K. Whynes. 2001. Willingness-to-pay for colorectal cancer screening. *European Journal of Cancer* **37** 1746–1751.

Fuloria, Prashant C., Stefanos A. Zenios. 2001. Outcomes-adjusted reimbursement in a health-care delivery system. *Management Science* **47**(6) 735751.

Gafni, Amiram, Stephen Birch. 2006. Incremental cost-effectiveness ratios (ICERs): The silence of the lambda. *Social Science & Medicine* **62** 2091–2100.

Gold, Marthe R., Joanna E. Siegel, Louise B. Russell, Milton C. Weinstein. 1996. *Cost-Effectiveness in Health and Medicine*. Oxford University Press, New York.

Hammerschmidt, Thomas, Hans-Peter Zeitler, Reiner Leidl. 2004. A utility-theoretic approach to the aggregation of willingness to paymeasured in decomposed scenarios: development and empirical test. *Health Economics* **13** 345–361.

Harvard Center for Cancer Prevention. 2008. *Tools and Strategies to Increase Colorectal Cancer Screening Rates: A practical guide for health insurance plans*. Harvard School of Public Health.

Hirth, Richard A., Michael E. Chernew, Edward Miller, A. Mark Fendrick, William G. Weissert. 2000. Willingness to pay for a quality-adjusted life year: In search of a standard. *Medical Decision Making* **20** 332–342.

Jack, William. 2005. Purchasing health care services from providers with unknown altruism. *Journal of Health Economics* **24** 7393.

Jonas, Daniel E., Louise B. Russell, Robert S. Sandler, Jon Chou, Michael Pignone. 2008. Value of patient time invested in the colonoscopy screening process: Time requirements for colonoscopy study. *Medical Decision Making* **28** 56–65.

Kahneman, Daniel, Amos Tversky. 1979. Prospect theory: An analysis of decision under risk. *Econometrica* **47** 263–291.

Keeney, Ralph L., Howard Raiffa. 1993. *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. Cambridge University Press, New York.

King, Joseph T., Joel Tsevat, Judith R. Lave, Mark R. Roberts. 2005. Willingness to pay for a quality-adjusted life year: Implications for societal health care resource allocation. *Medical Decision Making* **25** 667–677.

Klose, Thomas. 2003. A utility-theoreticmodel for QALYs and willingness to pay. *Health Economics* **12** 17–31.

Laffont, Jean-Jacques, David Martimort. 2001. *The theory of incentives: the principal-agent model*. Princeton University Press.

Lee, Chris P., Stefanos A. Zenios. 2008. A shadow price framework for quantifying health care demand, spending and disparity. Working Paper.

Levin, Bernard, David A. Lieberman, Beth McFarland, Robert A. Smith, et al. 2008. Screening and surveillance for the early detection of colorectal cancer and adenomatous polyps, 2008: A joint guideline from the American Cancer Society, the US Multi-Society Task Force on colorectal cancer, and the American college of radiology. *CA: A Cancer Journal for Clinicians* **58** 130–160.

Lewis, James D., David A. Asch. 1999. Barriers to office-based screening sigmoidoscopy: Does reimbursement cover costs? *Ann Intern Med* **130** 525–530.

Ma, Ching-To Albert. 1994. Health care payment systems: Cost and quality incentives. *Journal of Economics & Management Strategy* **3**(1) 93–112.

Ma, Ching-To Albert, Thomas G. McGuire. 1997. Optimal health insurance and provider payment. *The American Economic Review* **87**(4) 685–704.

Marioso, Begoa Garcia, Izabela Jelovac. 2003. Gps payment contracts and their referral practice. *Journal of Health Economics* **22** 617635.

Mas-Colell, Andreu, Michael D. Whinston, Jerry R. Green. 1995. *Microeconomic Theory*. Oxford University Press, New York.

McClellan, Mark. 1997. Hospital reimbursement incentives: An empirical analysis. *Journal of Economics & Management Strategy* **6**(1) 91128.

Miyamoto, John M., Peter P. Wakker, Han Bleichrodt, Hans J. M. Peters. 1998. The zero-condition: A simplifying assumption in QALY measurement and multiattribute utility. *Management Science* **44** 839–849.

Mooney, Gavin, Mandy Ryan. 1993. Agency in health care: getting beyond first principles. *Journal of Health Economics* **12** 125–135.

National Cancer Institute. 2007. *Cancer Trends Progress Report*. National Cancer Institute. Available via http://progressreport.cancer.gov/trends-glance.asp [accessed January 1, 2010].

Newhouse, Joseph P. 1996. Reimbursing health plans and health providers: Efficiency in production versus selection. *Journal of Economic Literature* **34**(3) 1236–1263.

Ozer, Ozalp, Wei Wei. 2006. Strategic commitments for an optimal capacity decision under asymmetric forecast information. *Management Science* **52**(8) 1238–1257.

Pasternack, Barry Alan. 1985. Optimal pricing and return policies for perishable commodities. *Marketing Science* **4**(2) 166–176.

Pauly, Mark V., Philip J. Held. 1990. Benign moral hazard and the cost-effectiveness analysis of insurance coverage. *Journal of Health Economics* **9** 447–461.

Pignone, Michael, Dawn Bucholtz, Russell Harris. 1999. Patient preferences for colorectal cancer screening. *Journal of General Internal Medicine* **14** 432–437.

Pignone, Michael, Melissa Rich, Steven M. Teutsch, Alfred O. Berg, Kathleen N. Lohr. 2002. Screening for colorectal cancer in adults at average risk: A summary of the evidence for the u.s. preventive services task force. *Annals of Internal Medicine* **137** 132–141.

Pliskin, Joseph S., Donald S. Shepard, Milton C. Weinstein. 1980. Utility functions for life years and health status. *Operations Research* **28** 206–224.

Polsky, Daniel. 2005. Does willingness to pay per quality-adjusted life year bring us closer to a useful decision rule for cost-effectiveness analysis? *Medical Decision Making* **25** 605–606.

Roberts, Stephen D., Lijun Wang, Robert L. Klein, Reid M. Ness, Robert S. Dittus. 2007. Development of a simulation model of colorectal cancer. *ACM Transactions on Modeling and Computer Simulation (TOMACS)* **18** 4:1–4:30.

Rosenthal, Meredith B. 2008. Beyond pay for performance - emerging models of provider-payment reform. *New England Journal of Medicine* **359**(12) 1197–1200.

Schrage, Linus. 2008. *Optimization Modeling with LINGO*. LINDO SYSTEMS INC, Chicago, IL, 6th ed.

Seror, Valerie. 2008. Fitting observed and theoretical choices womens choices about prenatal diagnosis of down syndrome. *Health Economics* **17** 557–577.

Shaheen, Nicholas J., David F. Ransohoff. 1999. Sigmoidoscopy costs and the limits of altruism. *American Journal of Medicine* **107** 286–287.

Shleifer, Andrei. 1985. A theory of yardstick competition. *The RAND Journal of Economics* **16**(3) 319–327.

Shumsky, Robert A., Edieal J. Pinker. 2003. Gatekeepers and referrals in services. *Management Science* **49**(7) 839–856.

Simon, Carl P., Lawrence Blume. 1994. *Mathematics for economists*. W. W. Norton & Company.

Smith, James E., Ralph L. Keeney. 2005. Your money or your life: A prescriptive model for health, safety, and consumption decisions. *Management Science* **51** 1309–1325.

Smith, Robert A., Vilma Cokkinides, Otis Webb Brawley. 2008. Cancer screening in the united states, 2008: A review of current american cancer society guidelines and cancer screening issues. *CA Cancer J Clin* **58** 161–179.

Swan, Judith, Nancy Breen, Ralph J. Coates, Barbara K. Rimer, Nancy C. Lee. 2003. Progress in cancer screening practices in the United States: Results from the 2000 national health interview survey. *Cancer* **97** 1528–1540.

Tafazzoli, Ali, Stephen D. Roberts, Reid M. Ness, Robert W. Klein, Robert S. Dittus. 2009. Probabilistic cost-effectiveness comparison of screening strategies for colorectal cancer. *ACM Transactions on Modeling and Computer Simulation* **9**(2) 6:1–6:29.

Tomlin, Brian. 2003. Capacity investments in supply chains: Sharing the gain rather than sharing the pain. *Manufacturing & Service Operations Management* **5**(4) 317–333.

Tversky, Amos, Daniel Kahneman. 1992. Advances in prospect theory: cumulative representation of uncertainty. *Journal of Risk and Uncertainty* **5** 297–323.

Wang, Yunzeng, Yigal Gerchak. 2003. Capacity games in assembly systems with uncertain demand. *Manufacturing & Service Operations Management* **5**(3) 252–267.

Ward, Elizabeth, Michael Halpern, Nicole Schrag, Vilma Cokkinides, et al. 2008. Association of insurance with cancer care utilization and outcomes. *CA A Cancer Journal for Clinicians* **58** 9–31.

Weng, Z. Kevin. 1995. Channel coordination and quantity discounts. *Management Science* **41**(9) 1509–1522.

Yaesoubi, Reza, Stephen D. Roberts. 2009. Coordinating contracts in a preventive health care system. Edward P. Fitts Department of Industrial and Systems Engineering, North Carolina State University, Raleigh, NC 27695-7906, U.S.A.

Zank, Horst. 2001. Cumulative prospect theory for parametric and multiattribute utilities. *Mathematics of Operations Research* **26** 67–81.